AD 651 781

PRINTED CLASSIFICATION TEST

J.P. Guilford, et al

Army Air Forces
Washington, D.C.

1947

NOTICE TO DEFENSE DOCUMENTATION CENTER USERS

This document is being distributed by the Clearinghouse for Federal Scientific and Technical Information, Department of Commerce, as a result of a recent agreement between the Department of Defense (DOD) and the Department of Commerce (DOC).

The Clearinghouse is distributing unclassified, unlimited documents which are or have been announced in the Technical Abstract Bulletin (TAB) of the Defense Documentation Center.

The price does not apply for registered users of the DDC services.

Army Air Forces
Aviation Psychology Program
Research Reports

Printed Classification Tests

REPORT NO. 5

Edited by
J. P. GUILFORD
Professor of Psychology
University of Southern California
With the Assistance of
JOHN I. LACEY
Research Associate
Samuel S. Fels Research Institute

1947

Preface

This volume attempts to present not only a definitive account of one aspect of a vast project in vocational-test development, but also a useful record of the experiences gained in the execution of that project. To the extent that it succeeds, it will be of value not only to aviation psychologists who carry on in the service of military or civilian authorities, but also to vocational psychologists in general. While the tone of the volume is pitched to the ear of the professional psychologist, an attempt has been made to avoid the more technical jargon of the more specialized statistically minded. By confining himself to the less technical passages, the lay reader may find much that is illuminating and interesting concerning tests and test methods.

Although there was no attempt, in the program to be described, to follow any preconceived ideal procedure of test development, inherent in this account is an emerging pattern of research, which, utilizing many of the techniques of the past, suggests what such a program can be when liberal support, in the form of trained personnel, suitable equipment, and an almost unlimited number of experimental subjects, is provided.

Well-known test theories, and past experiences in their application, were brought to bear upon the problems of vocational selection and classification in a rather special area, though it was an area of enormous scope from a psychological standpoint. While the theoretical problem and the empirical test of a procedure always had to be subordinated to the fulfilment of a pressing practical goal, there is, nevertheless, many a finding that transcends the immediate problem and its solution. The best example of this was the utilization of factorial theory and methods.

Factorial analysis, brought into use somewhat incidentally at first, became eventually the centralizing and guiding principle in connection with most printed-test development. It must be admitted that the factorial studies were neither as well planned nor as well executed as they would have been in a program that had centered around them from the very beginning. Only near the end of the four years' research did their full benefits become apparent. Two ambitious intercorrelation studies, planned in the early months of 1945, were not completed in time to be treated in this report. The results of earlier analyses are given liberal mention, however, and the description and evaluation of tests lean heavily, and it is believed rather effectively, upon appeals to factorial information.

Rather unique to vocational-test research, also, is the inclusion of analysis of job criteria by the factorial methods. It is believed that in this direction lies an economical, systematic, and dependable procedure for coverage of aptitudes and for fitting tests to vocations.

In the presentation of results, efforts have been made to facilitate perusal of the chapters by the reader by means of a uniform type of description of tests. This was not easy in view of the varied types of tests, the nonuniformity of data available, and the multiple authorship. Where efforts along this line have faltered, some of the monotony that may arise from repetitious uniformity may be thus relieved. Fortunately, there had been considerable uniformity and system in record keeping and record treatment, testifying to wise supervision and to cooperation among field units. Variations in procedure over a 4-year period and over a number of research units at different field stations were inevitable. Most regrettable of all are the few omissions of data which leave gaps that were impossible to fill.

Failures are recounted as well as successes, but false starts that never reached the stage of yielding results are best left unreported. Errors undoubtedly still remain undetected in places, in spite of diligent efforts to minimize their number and seriousness. Besides the editor and the assistant to the editor, Capt. John I. Lacey, who have read all chapters a number of times, Col. John C. Flanagan, Maj. Robert L. Thorndike, Capt. Lloyd G. Humphreys, and Technical Sgt. Paul C. Davis have read most of them. All have made valuable suggestions that have been incorporated. None should be held accountable for errors that still remain.

The editor has exercised considerably more than the usual editorial prerogatives, in that he has taken the liberty to suggest, and even to make, omissions, modifications, and additions in places for the sake of greater internal consistency and uniformity of treatment and for the sake of more complete coverage of points that could be brought out. From this point of view, the authors should not be held too strictly accountable for all statements of theory or of interpretations that appear under their names. While the editor is willing to assume responsibility for the publication of statements of opinion, this does not necessarily mean that he subscribes fully to all opinions offered.

This report and the work for which it stands are the product of many minds and hands—many more, indeed, than those whose names appear herein. Like other reports in this series, it represents a genuinely cooperative program. The writers of the chapters that follow have been, in general, substantial contributors to the execution of the program (though not the only substantial contributors), as the numerous footnotes will testify. Other footnotes will show that there were many other sources of test ideas and test construction. Unnamed are the numerous persons, civilian as well as military, who have added their contributions by administering, scoring, recording, calculating, and other activities. By way of exception, there will be mentioned here the names of some who cannot be cited adequately in footnotes but who should receive mention for special accomplishments. Two artists, Sgt. Fredrick H. Meise and Cpl. James B. Ferguson, designed illustrations for test items as well as those pictured in this report. Pfc. Leland D. Brokaw carried most of the responsibility for assembling the statistics concerning tests. Mrs. Jeanne E. Russell worked tirelessly on the preparation of the final manuscript as well as in keeping organized files on tests. Maj. Merrill F. Roff played an active role in the initial stages of much of the test-development program—much more than references in footnotes would indicate.

Footnote citations of credit for test development are given, first, to those who actually designed or wrote items; second, to those who contributed new test ideas; third, to those who criticized tests with significant consequences; and fourth, to those who supervised development in a significant manner. In the citations, contributors are named in alphabetical order, disregarding military rank and extent of contribution. Many of those who were present during the gestation and birth of a test have given their judgment as to the contributors who should be mentioned. In spite of great efforts to be just, many inequities will still be apparent to some. It is believed, however, that less injustice is done in terms of unwarranted inclusion or exclusion from a list of contributors, than would have been done in attempting to rank contributors for relative merits.

J. P. GUILFORD,
Colonel, Air Corps.

BEVERLY HILLS, CALIF., September 1946.

Contents

CHAPTER ONE

Job Requirements of Aircrew[1]

INTRODUCTION

Contents of the Chapter

The main purpose of this chapter is to describe very briefly the kind of men required for each of the three air-crew assignments—bombardier, navigator, and pilot. No space will be given to describing the duties of air-crew members, since adequate descriptions are given in other volumes of this series. It is sufficient here to give a synopsis of the information upon which were based the many ideas for tests accounted for in this volume.

The first section of the chapter will present a brief list of the sources of information concerning the psychological requirements of air-crew jobs. Three sections will give short descriptions of these requirements and their relative importance for each air-crew job. A final section states some very general considerations.

SOURCES OF INFORMATION

An examination of the list of sources of information concerning air crew reveals that many different approaches were made to job analyses. For a more adequate account of them, the reader is referred to Report Nos. 3, 8, 9, and 10. It is recognized that most of these procedures have their weaknesses, but since we are concerned here only with positive values, no criticisms will be offered.

Types of Information

The various types of information and their sources were as follows:

Faculty board proceedings.—When a student is eliminated from pilot training, his instructors and check pilots prepare a statement concerning (1) the student's personal traits, emphasizing deficiencies, and (2) the manner in which he flew his plane. Similar reports are also available in connection with bombardier and navigator training.

Flying evaluation board reports.—If an air-crew man who has earned his wings is found to be unsuited to tactical flying for any reason, his case is submitted to a local evaluation board. When the board has reached a decision, the report with recommendation is forwarded to a central board at Headquarters, Army Air Forces. The man is then either kept on flying status or is reclassified. In the report, statements regarding his experi-

[1] Written by Maj. William E. Walton.

ence, present assignment, attitude towards his job, and apparent proficiency in flying are included.

Observations of training.—From the first days of the AAF psychological program, aviation psychologists attempted to learn all that was possible about flying. Manuals, training memoranda, and textbooks were studied by selected personnel. Missions were flown with students. Officers and enlisted men were sent out under temporary duty orders to make studies of specific air-crew assignments or of some phase of the assignment. One example is an extended visit of an officer and several enlisted men at a primary pilot school [2] to make a study of the task of landing an airplane. Another visit was made to a bomber training school to observe the activities of crew members in training for combat operations. At later stages, rated officers—pilots, bombardiers, and navigators—were assigned to various psychological units or projects for extended periods of duty. Some of these had had some degree of professional psychological training.

Formal job analyses.—These analyses consisted of setting up checklist forms, similar to those used in industries, and making a fairly complete survey of men and their jobs, with special emphasis upon the psychological traits required. The reports of results included such topics as: general duties of pilot and commander of crew (in the case of a pilot analysis); nature of work, including location in airplane, posture, and working area; equipment and tools, including delicate, as well as gross, manual controls; computational aids, such as slide rules and tables; types of work required (described under sequence of duties); movements required, duration of work, and speed required; related vocations or avocations; responsibilities; job satisfactions; description of worker as to experiences, physical and mental abilities; and personal qualities, including interests and attitudes.

Interviews with eliminated cadets.—Realizing that there were weaknesses in the reports of faculty boards and flying-evaluation boards, an attempt was made to understand the job of the pilot or, to be more specific, of the student in pilot training by an interview approach.

Rating scales for aviation cadets.—Beginning early in 1942 a proficiency rating scale designed by aviation psychologists was used in all primary pilot schools. As contrasted to the faculty-board proceedings in which instructors stated in their own terminology why a student was eliminated, the rating scale carried a list of 20 traits, which thus provided a report in standardized terminology.

Ratings by students concerning difficulties experienced in learning to fly.—An interview rating scale containing 24 items was presented to students in basic pilot training. They were asked to indicate on a checklist

[2] The primary school provides the first stage of flying training for the pilot. This stage is sometimes, but rarely, referred to as elementary pilot training. Primary training is preceded by a pre-flight phase, which is composed entirely of ground-school courses, and is followed by the basic and advanced flying-school phases.

the extent to which they had found each item difficult in (1) primary training, and (2) basic training.

Grade-slip entries.—As a routine procedure, each pilot instructor made an entry on a grade slip indicating any difficulties or weaknesses that the student with low grades exhibited. These data were analyzed and categorized in more suitable form for a job-analysis study.

Clinical studies.—Several very ambitious clinical studies were made in an attempt to reveal fundamental personal characteristics that are important determiners of air-crew success. This entailed observing, interviewing, examining, and measuring the performances of individuals in training situations. In this connection, psychologists lived with students at flying schools, taking flying training with them, messing with them, and living in cadet quarters with them.

Anecdotal summaries.—In several instances, the anecdotal method was used in preparing reports on job analysis. Collection of instances believed to show good and poor judgment is one example of the use of the method.

Instructors' and supervisors' checklist data.—There have been several variations of this approach. In one, flying instructors merely ranked 20 items according to their importance; in another, they checked the important ones; and in the third, they rated each one according to a numerical scale. Average ranks, frequency of mention, and average ratings were used in the summaries.

In one extensive study in the Eighth, Ninth, Twelfth, and Fifteenth Air Forces, supervisors of air-crew personnel were asked to indicate the relative importance of each of 20 traits for individuals "capable of doing superior work of a specific type in combat operations." These officers indicated, on a 9-point rating scale, the minimum acceptable standards which they believed should be met for each of these traits in selecting and classifying air-crew personnel.

ANALYSIS OF THE BOMBARDIER'S JOB

Psychological Description

Of the many psychological characteristics required of the bombardier, perhaps the most important are the ability to attend to a variety of detailed activities and the ability to remember the serial order of events. The bombardier must be able to judge minimal rates of movement (rate and drift) and must be able to synchronize these movements. This calls not only for perceptual judgment, but precision of eye-hand coordination. He must be able to work calmly under pressure of time, and he must, therefore, be free from fear or nervousness. He must not be tense as he coordinates the movements of the knobs in killing rate and drift. He must be alert to his job, work rapidly, and make quick adaptations. He must be able to identify the target and orient himself spatially. These are some of the principal traits [3] demanded of the bombardier.

[3] In this Report the term "trait" will be used in a very general sense to include abilities (see ch. 22).

Relative Importance of Various Categories

Tables 1.1 and 1.2 show respectively: (1) A list of factors causing elimination from bombardier schools and percentages of times they were mentioned in elimination reports on 102 eliminees at 1 school and (2) ratings on a 9-point scale of 20 psychological categories, made by supervisors of combat teams.

Comparison of the two analyses.—The lists of traits in tables 1.1 and 1.2 are so different in terminology that it is perhaps futile to attempt to look for similarities and differences in degree. Were they more alike, one might well look to them to supply some information regarding the communality of requirements for training and for combat. Unfortunately, this comparison is limited to some very general observations. In the train-

TABLE 1.1.—*Frequencies of reasons for elimination from bombardier training as found in 102 elimination board reports in one school*

Categories	Code[1]	Percent times mentioned
Ability to execute a series of activities accurately and in proper order	C	70
Ability to learn	I	57
Eye-hand coordination	C	55
Ability to work rapidly	A	47
Ability to make fine and smooth manual movements	C	41
Nervousness and tenseness	P	37
Ability to adapt to unusual circumstances	P	28
Self-confidence	P	22
Judgment	P	11
Self-analysis	I	11
Interest and motivation	P	11
Ability to perform arithmetic computations accurately and rapidly	I	10
Orientation	A	8
Ability to perceive minimal movements	A	6

[1] The 16 items were grouped into 4 main categories, each with a code letter as follows: Intelligence and judgment (I), alertness and observation (A), coordination and technique (C), and personality and temperament (P).

TABLE 1.2.—*Average ratings of importance of psychological categories for combat bombardiers*[1]

Category	Mean rating	Category	Mean rating
Orientation and observation	7.8	Coordination	6.1
Emotional control	7.6	Motivation	6.1
Speed of decision and action	7.3	Leadership	6.0
Judgment	7.0	Arithmetic calculations	5.8
Finger dexterity	6.9	Estimation of speed and distance	5.8
Memory	6.8	Reading comprehension	5.7
Dial and table reading	6.8	Visualization of the flight course	5.6
Division of attention	6.8	Mechanical comprehension	5.4
Serial reaction time	6.5	Mathematics	5.2
Dependability	6.5	Arithmetic reasoning	5.1

[1] Raters were 48 squadron and group bombardiers. The ratings were made on a 1—9-point scale under the instruction "circle the number indicating the minimum standard which you believed should be required." Definitions of scale numbers were roughly as follows: 9—exceptional; 7—very much better than average; 5—better than average; 3—average enlisted man; 1—worse than average.

ing data, quick, smooth, and accurate motor coordinations are stressed as important; personality traits are moderately important; and perceptual and intellectual traits—including arithmetic calculations—are near the bottom of the list. In combat, on the other hand, intellectual and perceptual abilities seem to rate higher, though arithmetic calculations still are relatively low, and some personality traits increase in importance. On the whole, there is little agreement between the two lists. Whether there would be a closer agreement between two independent groups of judges either in training or in combat activities is unknown.[4]

ANALYSIS OF THE NAVIGATOR'S JOB

Psychological Description

It has been said that the navigator is the most intellectual of the air crew, that he is pedagogically inclined, and academically motivated. Every analysis of his duties has emphasized the high degree of mentality required by this position. Whether or not the good navigator need be a more intelligent person than the good pilot or good bombardier can be questioned. The matter cannot be settled without defining "intellectual" and "intelligent" in some demonstrable terms.

There is no doubt that in certain abilities the navigator must excel. The very nature of his work demands that he be interested in and have some knowledge of mathematics, though this need not include "higher" mathematics. It is certain also that he must readily understand abstract concepts. As we examine reports of eliminated cadets and job descriptions prepared by instructors in navigation, we are impressed with the large number of other traits needed by the navigator. These include such traits as the ability to work rapidly, accurately, and neatly. With respect to the last-mentioned trait, it is a fact that a number of students have been eliminated because they were either poor draftsmen or could not write legibly enough to read their own figures while making computations in the air. In other instances serious errors have been made in the navigator's log books for a dozen or more reasons, not the least of which were errors in simple addition and subtraction.

The navigator must also be thorough in his work and able to analyze and to correct his own errors. He must exercise good judgment and show the ability to concentrate effectively on navigational problems over long periods of time. Some individuals have been eliminated because they failed to precheck their instruments and others because they failed to report defective instruments upon landing. While manual skills are perhaps not so important as intellectual abilities, we do find that some navigators have difficulty in manipulating such instruments as the drift meter and the pelorus or astrocompass.

The navigator's confidence in his work must be a balanced mental trait. This means that he must have neither too much nor too little confidence.

[4] A more complete account of the bombardier will be found in report No. 9 of this series.

701120—47—3

He must not be so over confident that he takes only one reading with the satisfaction that it is correct. He has been taught that frequent readings minimize the probability, as well as the size, of errors that can be made. He must not be so lacking in confidence that he takes an excessive number of readings. By so doing, he passes on his apparent lack of confidence to the rest of the crew members who may consequently suffer lowered morale. The navigator must display coolness and deliberation. This is especially true in combat, when, after the bombing run, the navigator must keep oriented in the midst of battle.

Another important characteristic of the good navigator is foresight and planning. In one combat mission, a forced landing was necessary. When the pilot called his navigator for suggestions as to an emergency landing field, he was immediately told that a few miles further on there was a beach on which the navigator thought it might be possible to make a crash landing. Later it was learned that it was the practice of this navigator to note all level fields and beaches that looked favorable for a crash landing and to mark them on his map along the course of flight for future reference.

Navigators are constantly impressed with the necessity of being familiar with several forms of navigation. If the navigator is flying CAVU most of the time, he may neglect to keep up on dead-reckoning procedures. Finally, the navigator must be a leader of men, because he is usually considered third in command of the ship.

Relative Importance of Navigator Qualities

Table 1.3 shows some of the items commonly checked by instructors as causes for elimination from navigation training, while table 1.4 presents combat data comparable to those previously given in table 1.2 for the bombardier.

Comparison of the two analyses.—In training, arithmetic computation —in terms of both speed and accuracy—ranks very high. Judgment, visualization, reasoning, and ability to learn abstract concepts are also regarded as very important. Among temperamental traits, neatness and orderliness are deemed significant. Other personality traits are of moderate or low importance, and motor coordination ability is not mentioned at all.

In combat, certain temperamental traits come up to the head of the list, equaling or excelling intellectual traits, such as judgment and arithmetical computations which are still prized. Perceptual qualities are of moderate or low significance in training, but a perceptual trait—orientation and observation—heads the list for combat performance. Motor coordination is at the bottom of the list as judged by combat supervisors, in good agreement with opinion of instructors in training schools.[8]

[8] A much fuller account of the navigator will be found in report No. 10 of this series.

TABLE 1.3.—*Percentage of times traits were checked by 112 instructors as cause of elimination from navigation school*

Category	Percentage
Inability to correct own errors	76
Errors in simple arithmetic computations	75
Slowness in learning new concepts	74
Poor judgment	73
Slowness in simple arithmetic computations	72
Incapable of adequate visualization to perform celestial work	68
Inability to meet and adjust effectively to new situations (especially in the air)	68
Lack of analytical mind (regardless of mathematical training)	59
Lack of orderliness in work procedures or log book	56
Lack of confidence	49
Inability to concentrate effectively over prolonged periods of time (examinations and flights)	48
Nervousness in examinations	41
Lack of initiative	41
Nervousness in flights	39
Lack of neatness in chart work and log book	37
Inability to use computer	33
Inadequate mathematical background	30
Lack of necessary emotional stability	30
Inability to read drift	26
Inability to use tables or graphs	23
Lack of interest	23
Inability to shoot with sextant[1]	22
Inadequate general educational background	19
Inability to learn necessary technical terms	17
Inability to read or use instruments	14
Airsickness (as a contributing factor)	8
Inability or unwillingness to accept new concepts or techniques	7
Dislike of flying	4
Fear of flying	2

[1] Based on elimination in celestial navigation only.

ANALYSIS OF THE PILOT'S JOB

Psychological Description

In general, the pilot must be a person who thinks and acts in a quick and positive manner. This is perhaps more true of the fighter pilot than of the bomber pilot, who can at times be more deliberate in his thinking. A similar difference exists between fighter and bomber pilots with regard to speed of action. The latter's actions should be highly characterized by reliability and dependability.

Both types of pilot should show good judgment, although that of the bomber pilot is expected to be more mature. It is important for both men to remember procedures. In the case of the bomber pilot there are a few more things to do, and the order in which they are done is of great importance. The fighter pilot must be far more alert to what is going on around him than the bomber pilot, because the latter can depend upon

TABLE 1.4.—*Average ratings of importance of psychological categories for combat navigators*[1]

Category	Mean rating
Orientation and observation	7.8
Emotional control	7.3
Dependability	7.2
Judgment	7.1
Speed of decision and action	7.1
Reading comprehension	7.0
Arithmetic calculations	6.9
Memory	6.8
Division of attention	6.8
Dial and table reading	6.6
Estimation of speed and distance	6.6
Leadership	6.6
Motivation	6.5
Visualization of the flight course	6.4
Arithmetic reasoning	5.9
Serial reaction time	5.9
Mathematics	5.5
Mechanical comprehension	5.3
Finger dexterity	5.0
Coordination	4.8

[1] See footnote to table 1.2. The raters were 77 squadron and group navigators.

his many crew members to inform him of the presence and activity of enemy airplanes.

Differences between the two types of pilots are more apparent in temperamental traits than in abilities. The good fighter pilot should be an aggressive individual but, in that aggressiveness, should not lose control of his emotions. A trait common to both is the ability to work in a team. The bomber pilot must inspire his crew, give them a feeling of confidence in him and in his decisions, and develop in them a spirit of cooperation. He is expected to develop a comradeship with his crew without permitting the element of familiarity to destroy his discipline. The fighter pilot does not always function as a "lone eagle" in his combat operations. He must frequently cooperate with others.

In addition, the average pilot must have, at least to a moderate degree, abilities ascribed to the navigator. He must possess ability to orient himself quickly and to match geographical landmarks with their representations on a map. Some pilots, particularly fighter pilots, must also possess characteristics of a good gunner, since they may be flying pursuit ships and engaging in either air-to-air firing or in strafing activities.

Relative Importance of Traits of Pilots

Tables 1.5 through 1.7 show the relative importance of various psychological categories as based upon elimination records, statements of eliminated cadets, statements concerning reclassified pilots, and judgments of supervisors of combat teams. Since most of the job-analysis work has been done upon the pilot, there are many such tables available, but those presented here will suffice to indicate some of the more important findings and some of the weaknesses in the job analyses which have been done. In table 1.5, five studies have been summarized. This summarization was possible because the same categories and method of evaluation had been used. The categories presented here have become somewhat standard in the AAF Aviation Psychology program.

Table 1.6 presents evidence concerning combat requirements. Supervisors of combat teams and others were asked to indicate on a 9-point rating scale the extent to which each psychological factor is important to a pilot. From these data can be obtained some conception of what the supervisors think are the important traits for fighter and bomber pilots.

Table 1.7 shows the percentage of times various reasons were mentioned by 150 eliminated pilots as reasons for elimination.

Comparison of different analyses.—An examination of tables 1.5 through 1.7 will show that results concerning traits regarded as important for the pilot depend upon a number of factors: (1) Whether training or combat is the test of proficiency; (2) stage of training; (3) type of airplane; and (4) whether judgment is made by boards, instructors, or by students.

In primary training, the leading traits as indicated in elimination-board proceedings are judgment, coordination, progress in developing skills,

TABLE 1.5.—*Percentage of times categories were mentioned as a cause of elimination or reclassification in pilot training*[1]

Categories	Elementary eliminations		Advanced eliminations		Operational[2] reclassifications	
	N = 1,000	N = 1,000	Single engine N = 100	Twin engine N = 100	N = 100	N = 100
A. Intelligence and judgment	[3]*68*	..	*96*	*83*	..	*22*
Judgment	50	52	84	65	12	17
Foresight and planning	38	43	47	23	1	1
Memory	24	39	52	38	1	3
Comprehension	17	15	25	27	1	7
B. Alertness and observation	*70*	..	*89*	*75*	..	*7*
Visualization of flight course	36	30	46	41	1	0
Estimation of speed and distance	30	31	27	38	5	1
Sense of sustentation	24	34	25	7	3	0
Division of attention	28	41	43	14	3	3
Orientation	13	15	6	9	0	2
Speed of decision and reaction	15	39	35	40	7	2
C. Coordination and techniques	*81*	..	*91*	*69*	..	*42*
Coordination	58	36	74	37	0	0
Appropriateness of controls used	21	18	10	13	0	0
Feel of controls	2	37	25	3	0	0
Smoothness of control movement	22	25	30	25	0	1
Progress in developing technique	54	42	52	34	47	41
D. Personality and temperament	*43*	..	*50*	*67*	..	*91*
Absence of tenseness	22	29	28	30	1	12
Absence of confusion and nervousness	12	26	19	37	7	9
Absence of fear and apprehension	18	17	7	18	57	69
Suitable temperament	9	11	8	15	11	21
Motivation and attitudes	6	11	11	10	20	32

[1] Percentages do not total 100, since more than one factor is frequently given for each elimination.

[2] A very small percentage of these were actually in combat.

[3] Percentages in italics refer to relative frequencies with which *groups* of traits were mentioned.

foresight and planning, visualization of flight course, estimation of speed and distance, and division of attention. In advanced training, eliminations are most frequently said to occur in conjunction with deficiencies in judgment, coordination, memory, visualization, and progress in developing skills. This list differs chiefly from that for primary training in the addition of memory and the loss of foresight and planning. There are some differences between single-engine and twin-engine training, but they are of uncertain significance. In operational training, reclassified pilots most frequently show these characteristics: fear and apprehension, lack of progress, lack of motivation, and lack of judgment. The chief new feature, then, and it heads the list, is fear and apprehension.

TABLE 1.6.—*Average ratings of importance of psychological categories for combat pilot positions*[1]

Categories	Ratings by supervisors of combat teams: Fighter pilot	Ratings by supervisors of combat teams: Bomber pilot
Speed of decisions and reaction	8.0	7.2
Judgment	7.7	7.3
Motivation	7.7	6.4
Emotional control	7.6	7.3
Estimation of speed and distances	7.5	6.1
Division of attention	7.5	6.8
Leadership	7.4	5.9
Dependability	7.2	6.5
Orientation and observation	7.2	5.5
Visualization of the flight course	6.7	6.4
Memory	6.6	6.4
Coordination	6.1	6.0
Mechanical comprehension	6.0	6.0
Serial reaction time	5.9	5.9
Reading comprehension	5.6	5.7
Arithmetic reasoning	4.8	4.7
Dial and table reading	4.8	5.6
Finger dexterity	4.2	5.0
Arithmetic calculations	4.1	4.5
Mathematics	3.3	3.9

[1] Raters of fighter-pilot requirements were 30 squadron commanders and squadron operations officers in the European theatre of operations. Raters of bomber-pilot requirements were 117 similar officials.

TABLE 1.7.—*Percentage of times categories were mentioned by 150 eliminated cadets as cause of elimination (Pilot)*

Categories	Percentage of time mentioned	Categories	Percentage of time mentioned
Nervousness in the air	54	Understanding of plane's behavior	16
Slow progress	53	Judgment	15
Judgment of height-speed in landing	30	Motor coordination	13
Lack of "feel of the ship"	27	Erratic performance	13
Inappropriate attitudes	26	Flight planning and pattern	13
Poor control of the ship in landing	23	Mechanical flying	9
Instructional problems	20	Inadequate correction for wind	9
Attention	19	Poor acrobatics	9
Stick and rudder control	17		

In combat, traits rated among the most important for both fighter and bomber pilot are: Judgment, motivation, speed of decision and reaction, emotional control, and division of attention. Speed of decision and reaction is apparently much more crucial in combat than in training, as one might expect. It is interesting that whereas estimation of speed and distance is given high place for the fighter pilot, dependability is regarded more important for the bomber pilot.

Of all traits, judgment stands out as being most persistent and universal. This is not the place to try to define judgment or to break it down psychologically. In the minds of aviation observers it undoubtedly means a great variety of things. At best, it signified good or bad decisions (where "good" and "bad" mean that the result turned out well or did not turn out well, or that the decision was or was not what the observer would have done under similar circumstances). However this may be, the frequent mention of judgment for the pilot, and for other air-crew personnel as well, was a persistent challenge to break it down to manageable components and to devise tests for it.*

CONCLUDING STATEMENT

During the early months of the war, at least, job-analysis information from all known sources was eagerly grasped and exploited for what it seemed to be worth, in accordance with the desperateness of the situation. It was recognized that much better knowledge was needed and would probably be forthcoming during the later course of events. From the early days, when even anecdotal material was tolerated, and informal observations served as a basis for test ideas, the progress in job analysis was marked by a transition through statistical studies of quasi-standardized observations, until at later times factor-analysis methods were invoked to study job criteria as well as tests. Since the latter type of results can be discussed only in connection with tests, and these need to be described, an account of such results will be reserved for later pages (see ch. 28). It will be seen during the course of succeeding chapters how well, and at times how poorly, observations of jobs yielded useful concepts and led to tests which did or did not measure significant aspects of air-crew aptitude.

* For a fuller account of the pilot see report No. 8 of this series.

CHAPTER TWO

The Program of Printed Test Development[1]

JOB ANALYSIS IN RELATION TO THE CONSTRUCTION OF PRINTED TESTS

The previous chapter discussed various sources of information about pilots, navigators, and bombardiers that were available to guide test construction. This chapter, which discusses the printed-test research program, starts with the relationship of job-analysis findings to test construction. For this purpose, it is convenient to distinguish two levels of job-analysis information.

Levels of Job-Analysis Information

Practically all job descriptions can be placed in two categories. Some do not go beyond a description of what the worker does. Descriptions of this sort might legitimately be termed "phenotypic" descriptions. They are most likely to lead to job-sample tests. In thinking of the job of the pilot, for example, some task involving a stick and rudder bar is immediately suggested. Other job descriptions attempt to describe the abilities used by the worker in his job. Such descriptions are more taxing psychologically; i. e., they are at a more profound level. They might, therefore, be termed "genotypic" descriptions. They are likely to lead to tests of functions or factors.

Phenotypic descriptions and work-sample tests.—It is a psychological truism that maximum validity for a single test for any criterion can usually be obtained by means of a work-sample test. The reasons for this are not hard to find. The work-sample test, insofar as it is a true sample of the job, will contain the valid factors in proportion to their proper weighting and will be on the average about as reliable as the criterion. It seems obvious that this procedure will be most successful for relatively simple criteria.

If the job is very complex, on the other hand, phenotypic job descriptions lead to tests sampling segments of the job. If table reading is involved, a table-reading test is constructed; map-reading activity suggests a test of map reading, etc. When such tests have been constructed, however, the usual finding is that their correlations with each other are high, so that the multiple correlation derived by combining several such tests will be little higher than the single highest validity coefficient in the group.

[1] Written by Capt. Lloyd G. Humphreys.

Work-sample tests, in addition, are not widely useful since they are "tailored" for a particular criterion. While these tests have not been overlooked completely, it is certainly true that they have not constituted a major emphasis in the test research reported in this volume.

Genotypic descriptions and tests of functions.—The use of genotypic job descriptions has been limited by lack of knowledge concerning human traits and their measurement. Once these traits have been defined—and the factor-analysis technique gives promise of greatly facilitating this step—tests can be constructed to measure the separate functions. Although considerable progress had been made in this direction, chiefly due to the work of Thurstone, a satisfactory battery of tests of independent functions or factors was not in existence at the outset of printed-test construction in the Army Air Forces Aviation Psychology Program.

The advantages accruing through the use of tests of independent functions are substantial, particularly in a classification battery where a test may be weighted for more than one specialty. Such tests are also more flexible if criteria change. From the first, therefore, test research was oriented toward tests of important functions. Certain functions were deemed to be important in early job analyses. As validation studies of classification and experimental tests became available, the list of important functions was considerably modified and somewhat enlarged.

Available Job Information

For reasons discussed in the following section, the problem of selecting and classifying the pilot more or less dominated the research program from the first. Concerning the pilot, the most important source of job information available at the beginning of research with printed tests was the analysis of faculty board proceedings discussed in chapter 1. Comments made by flying instructors concerning reasons for elimination of 1,000 students in elementary flying training constituted the basic data. Psychological analysis of these comments produced a list of 20 traits that were presumably important in pilot success. No matter how keen the analyst, any analysis of comments made by psychologically untrained observers would be deficient, because the basic data are not completely sound. Although this was realized from the outset, this list of 20 traits constituted almost all the information available concerning the abilities necessary in learning to fly "the Army way." It should be noted that this list oriented the research program from the beginning towards tests of functions or factors.

THE PLAN OF TEST DEVELOPMENT

Importance of the Analysis of Faculty Board Proceedings

Although faculty-board proceedings had been studied only for pilots, the organization of the research program, as well as the planning for printed-test research, was based on the analysis of those proceedings. This was the result of several circumstances. In the first place, the original responsibility of the aviation psychology program was for research on pilot selection; responsibility for bombardiers and navigators was assumed somewhat later. In the second place, pilot quotas were initially so large in comparison to those for navigators and bombardiers that the classification problem was largely a pilot-selection problem. In addition, a satisfactory degree of validity was obtained very early for the navigator aggregate aptitude score, while the available bombardier criterion had so little reliability that research concerning bombardier aptitude was almost hopeless.

Organization of the research program.—The list of 20 traits derived from the study of elimination board proceedings was divided into four main categories: Intellectual, perceptual, temperamental, and psychomotor. Responsibility for test research was originally delegated as follows: Psychological Research Unit No. 1, temperament tests; Psychological Research Unit No. 2 and the Department of Psychology of the School of Aviation Medicine, psychomotor tests; Psychological Research Unit No. 3, intellectual and achievement tests; and the Psychological Section, Headquarters, AAF Training Command, perceptual tests. While the responsibility for test development in these areas was later modified in several ways, the separation of tests into these categories continued to be a factor in test development until the end of the program. It should be noted that, since the concern of the present volume is with printed tests, only three of the four categories will be discussed. Psychomotor tests constitute the group of apparatus tests discussed in Report No. 4 of this series.

The test coding system.—The coding system established for the test-research program was based upon the same four categories. The 20 hypothesized traits of unsuccessful pilots made up most of the subcategories used in the system.[2] The basic code number for a test begins with two letters followed by three digits and then another letter. All classification tests, or tests designed for classification purposes, have code numbers beginning with the letter "C." The second letter indicates one of the four main categories: I—Intellectual; P—Perceptual; E—Temperamental; and M—Pyschomotor. The first digit indicates the subarea within the main area. The next two digits indicate different tests within the subarea. The following letter indicates different revised forms of the same test. This basic code number is followed, in the case of tests in other than final form, by the letter "X." Successive experimental versions of the same form, therefore, are indicated as X1, X2, etc. Thus, the code number CI206C (Arithmetic Reasoning) means that the test was designed for classification purposes, in the intellectual area, reasoning subgroup, and that it was the third form of the sixth reasoning test to be given a code number.

Plan of research.—The original plan of research was to develop one or more tests in each of the subcategories of the coding system for vali-

[2] Exceptions will be discussed in later chapters.

dation and possible inclusion in the classification battery. This procedure was not, of course, deemed to be a permanent solution to the pilot-selection problem. It did promise to give initial coverage of a number of potentially valid factors. It was expected that validation findings and additional job analyses of various types would serve as the primary guide in later research.

Importance of Validation Studies

The importance of rapid validation of tests cannot be over-emphasized, either in the research program in the Army Air Forces or in any selection program. The usefulness of any job analysis and subsequent test construction is determined by the correlations of the tests with criteria. Knowledge of the criteria used is necessary, therefore, in order to evaluate the statistics concerning individual tests to be reported in the chapters to follow.

The pilot criterion.—The criterion of success as a pilot routinely used in validation studies was graduation or elimination from primary flight training. Most eliminations usually occurred during primary training.[3] A smaller proportion of students was eliminated from basic training and a still smaller proportion from advanced and transitional. In all three phases the great majority of eliminations was for flying deficiency.[4] Few eliminations from pilot training for academic deficiency occurred either in the ground-school phase of flying training or in the preflight school.

After a student was classified, he spent 2 months each in preflight, primary, basic, and advanced training. Using the criterion of eliminations, in primary training, validity data matured in a minimum period of from 2 to 5 months depending on when a test was given. When a classification-battery test was to be validated, a period of approximately 5 months was required. Many experimental tests were also given during the classification period so that the same time lag existed for them. Other experimental tests were given to classified pilots as they finished preflight training. Data on these men were then available in 2 months. This procedure made possible quick validation of many experimental tests.

The navigator criterion.—The standard criterion of success as a navigator was graduation or elimination from advanced navigation training, the only navigation phase of training beyond preflight. The important variables entering into this criterion were few in number. These were grades in navigation theory, ground missions, and flight missions, of which the third was most heavily weighted.[5] Every evidence indicates that this criterion was quite reliable.

Because of the small proportion of students classified as navigators, validation analyses for navigation were almost restricted to classification

---

[3] Basic eliminations exceeded primary eliminations for a few months in one of the three flying training commands. For this reason, certain tests were revalidated against the criterion of graduation-elimination through basic training.

[4] Evidence is available to show that flying deficiency means much the same thing during all stages of training. Pilot criteria are discussed fully and critically in Report No. 8 of this series.

[5] The navigation criterion is discussed more fully in Report No. 10 of this series.

tests. Many months were necessary to accumulate as many as 1,000 cases of classified navigators on a test given during the classification period at a single classification center. With time in preflight and advanced navigation added, validation on a sufficiently large sample of a test for the navigator criterion took approximately one year. It later became possible to test a few classes of classified navigators graduating from preflight in all three flying training commands with small batteries of experimental tests.

The bombardier criterion.—The successful bombardier, for validation purposes, was the graduate from advanced bombardier school. Graduation or elimination was largely determined, in turn, by the "average-circular-error" and "percent-hits" scores obtained on practice bombing missions. The instructor's judgment concerning a student's capability as a bombardier also entered into the decision to graduate or eliminate, but in a nonsystematic fashion. Since the objective measures of bombardier proficiency, i. e., circular error and percent hits during individual training, are known to have had practically zero reliability, any reliability in the graduation-elimination criterion was probably due to the subjective judgments of instructors. That the bombardier criterion did have some degree of reliability is shown by the consistent positive correlations obtained between certain tests and that criterion.*

The same comments made concerning the relatively small number of classified navigators also apply to bombardiers. Adequate samples were difficult to obtain on tests other than those in the classification battery until a few classes of preflight graduates were tested with small batteries of experimental tests. The problem was made even more complicated by the unreliability of the criterion. If the top possible correlation between a test and a criterion is, for example, 0.30, one cannot be reasonably certain that any correlation at all exists unless very large numbers of cases are available.

Test Construction by Subareas

In order to carry out the plan to construct and validate at least one test in each subarea, the problem immediately arose as to when a test did or did not measure any hypothesized ability. The first step is an obvious one. If one cannot be certain that a given test is a good measure of the ability, a number of tests should be constructed in the subarea and experimentally administered. In selecting representative tests of the ability, reliability is a possible criterion. Within rather wide limits, however, reliability was considered to be relatively unimportant. Much more important were the intercorrelations of the experimental tests and their correlations with tests then in the classification battery. The technique of factor analysis, which is best described as an extension of correlational analysis, was therefore considered to be an important aid in selecting tests to measure the ability in question.

* For a more complete discussion of the bombardier criterion, see Report No. 9 of this series.

USE OF CORRELATIONAL AND FACTOR ANALYSES IN TEST CONSTRUCTION

Determining Uniqueness of Contribution

It is becoming increasingly evident that, in addition to the concepts of reliability and validity, the concept of uniqueness of contribution or purity deserves a central place in test construction theory. When one is faced with the practical problem of putting together a battery of tests to predict a criterion, individual test reliabilities and validities shrink in importance. Beta weights, which are a function of test intercorrelations as well as validities, become the criteria on which a test is accepted or rejected.

Relationship to correlational analysis.—If a test contributes information concerning individual differences over and above that furnished by a battery of other tests, that fact can be ascertained through correlational analysis alone. The multiple correlation between the test and a reference battery, when corrected for attenuation, must differ significantly from 1 if the test is to make a real contribution. This contribution consists of the measurement of a new function or functions.

Relationship to factor analysis.—Correlational analysis alone is sufficient to assess a test's contribution to a battery. Factor analysis is necessary in order to define the nature of that contribution. While the objectivity of the application of factor analysis to this and similar problems may have been overrated, the usefulness of the technique definitely has not. Factor results constitute an indispensable aid to the test constructor who is interested in what his tests measure and why they are valid. One very important use is to gain insight into the functions responsible for beta weights in regression equations.

In deciding which tests in a group designed to measure "foresight and planning," for example, were most worth validating, factor analysis was a considerable aid. A supposed foresight-and-planning test may, for example, turn out to be functionally very like the Arithmetic Reasoning Test already in the classification battery. No matter how different the apparent content of the two tests may be, the experimental test could not have a high priority for validation. A second foresight-and-planning test, on the other hand, may reliably define a new factor. Whether or not the new factor should now be given the name of the hypothetical function it was designed to measure is not always determinable. The test which best measures the factor, however, should certainly be validated.

A Guide to Test Construction

In a previous section, it was stated that validation findings were expected to guide test construction beyond the initial stages that resulted from the available information concerning the jobs of the pilot, navigator, and bombardier. This turned out to be only partially true. Validation of a relatively few tests will usually be a sufficient guide to the construction of other valid tests. Beta weights of the additional tests, however, may not differ significantly from zero. For this reason, correlational and factor analysis became as important as validation findings in the guidance of test construction.

Increasing unique contribution.—Factor analysis serves a very useful function in pointing out the ways in which the unique contribution of a given test can be increased. For example, a new factor is discovered in an analysis on which no test has a loading greater than 0.40. The test with the highest loading on the factor also has high loadings on the verbal and numerical factors. The first step is to form an hypothesis concerning the nature of the new factor. Equally important is to decide what features of the test contribute toward the verbal and numerical loadings. The second step is to vary the content of the test, the directions, the method used in recording the answers, or the time limit so that the verbal and numerical loadings will be decreased and the loading on the new factor will be maximized. The new test is then administered along with selected reference tests in order to check the factorial make-up of the revision. This process may be continued until satisfactory results are obtained.

The need for new test construction is indicated by factor-analysis findings in yet another way. A test with good reliability may show very little communality with the rest of a test battery. It is relatively easy in most cases to convert a nonerror specific factor to a common factor by appropriate test construction. This is particularly important if the test is known to have validity for some specialty over and above that predictable from its known common-factor content. In this connection, it should be pointed out that the prediction of test validities on the basis of a summation of products of test loadings and criterion loadings on known factors has been quite successful. The evidence for this will be discussed in considerable detail in chapter 28.

Empirically derived categories.—Factor analysis promises to furnish the test constructor empirically-derived, orthogonal categories for his tests. Considerable progress has been made in establishing these categories, both by civilian and military psychologists. Empirically-derived categories are most useful to the test constructor in conjunction with job analyses. The job analyst, in using factor results, has a framework for his description of the job. The factor categories, in addition, direct the analyst's observations toward details of the job that might easily go unnoticed otherwise.

Scoring Formulae in Relation to Factor Findings

There are several approaches to the development and use of scoring formulae for tests. All are represented in the tests discussed in this volume. The final practice, which is recommended grew out of correlational and factor studies.

A priori formulae.—On the basis of the [illegible]ndom-guessing hypothesis, the probability of obtaining a correct answer by chance is 0.5 in a 2-choice test, 0.33 in a 3-choice test, etc. The formula $R-W/(k-1)$, where k is the number of alternative responses, is expected to convert a chance score to zero. A formula of this type has been very commonly used, even in power[7] tests, although it may have little empirical justification. This practice has one nonstatistical advantage—both examinees and psychologically unsophisticated critics can be told that even though the right answer can be guessed in a multiple-choice test, guessing will not be profitable.

Maximum-reliability formulae.—Right and wrong scores on a test can be weighted so that a maximum degree of reliability is obtained. This should not be done, however, unless it is known that right and wrong responses are both measuring the same thing. If the factor patterns of rights and wrongs are identical, then the maximum-reliability scoring formula will be identical with the maximum-validity formula discussed in the next section.

Maximum-validity formulae.—When a test is being considered in isolation, a maximum-validity formula will be found to be most useful. The formula which maximizes the correlation between a test and a given criterion may not be the same, however, for a different criterion. It is conceivable that a test would have as many scoring formulae as there are criteria that it is used to predict, if right and wrong scores actually measure different functions. Right and wrong scores are very likely to be factorially dissimilar, as a matter of fact, in any speeded test. A number of cases will be presented in the chapters to follow in which this is true.

Use of right and wrong scores separately.—The finding that rights and wrongs often measure different functions came late in printed-test research. As a result, the procedure that is now recommended has been followed in relatively few test analyses. It now seems clear that the best way to handle right and wrong scores is to treat them as separate variables in test validation and analysis. A scoring formula should not be used in a classification battery except in rare cases because beta weights for rights and wrongs may differ from one criterion to another. Retention and weighting of either score in the final battery should depend upon the respective beta weights determined from the matrix of intercorrelations[8] of the entire battery.

TYPICAL HISTORY OF A TEST

The following outline of the typical developmental history of a test does not cover all tests in this volume. It applies to intellectual and perceptual tests much more than to temperament tests. It is perhaps more ideal

[7] In a power test, unless there is an unusual number of omissions, the correlation between rights and wrongs will be [illegible] high that a scoring formula of any kind cannot be justified on any empirical basis.

[8] In computing beta weights, the data for both rights and wrongs should not be obtained from the test as a whole. Correlation of errors between rights and wrongs can be avoided by using scores from separately timed, and comparable, parts of the test.

than typical, but the writer is convinced that if the better aspects of this general procedure had been followed more religiously, test research would have been even more productive.

Choice of the Function to be Investigated

In the early days of printed-test construction, subareas of research were determined largely by the analysis of faculty board proceedings which was in turn reflected in the coding system. A representative test or tests in each subarea was desired. Higher headquarters indicated the order in which these tests were to be supplied. A good account of the development of tests in a subarea according to this plan is given in chapter 9, Foresight and Planning Tests.

Later, the decision to investigate a given function frequently arose from a combination of validation and factor-analysis findings. An example of this sort is found in chapter 10, Integration Tests.

Test ideas.—After the function to be investigated, e. g., foresight and planning, had been selected, the personnel assigned to test construction spent a period of time reading available job descriptions, interviewing flying personnel, discussing the problem among themselves, etc. Anything that might lead to a likely test idea was investigated.

As test ideas were originated—and they often multiplied in a remarkable fashion—those responsible were asked to enlarge upon them, to write tentative directions, to outline a few items, and to suggest the conditions for administration. At this stage, a weeding process was required. In the absence of the completed test, and therefore any data, this process had to be based upon professional judgment alone. Rarely, however, was the selection of an idea for further development the result of only one individual's judgment. In most cases, and ideally, this was the result of joint discussion. The chief criterion was the possibility of unique contribution. Potential reliability, testing time, adaptability to IBM answer sheets, and "face validity"[9] were other criteria used.

The available test ideas in a restricted area were thus reduced to a number, such as eight, that could be easily administered together for intercorrelational purposes. The planning of work on the selected tests was oriented from the start, therefore, toward bringing the entire group to completion at approximately the same time.

Item Writing and Criticism

There is an old saying that two heads are better than one. Experience has shown that this is true in test construction. Whenever possible, two men were assigned to the development of a single test, one with primary responsibility, the other with immediate supervisory functions possibly including one or two other similar tests. These two, working closely together, produced the experimental version of the test.

[9] Face validity refers to the characteristic of a test that makes it appear to have validity to unsophisticated observers. No test constructor seriously believed that face validity would add a significant increment to actual validity. The morale and good feeling of the examinee were considered to be sufficiently important to warrant some effort in this direction. The silencing of potential critics lacking in psychological sophistication was also a consideration.

703320—47—3

Before a test was produced for experimental administration, it was gone over carefully by several independent critics. This step involved more than mere copy reading. Fundamental conceptions of the character of the test were frequently questioned. The joint contribution of several capable individuals more often than not was superior to what any one alone could produce.

Experimental Administration and Item Analysis

Wherever possible, experimental forms of a test were administered in advance of the proposed correlational study. This was done in order to check the clarity of the directions, other problems of administration, and internal consistency. Tests with relatively complicated directions, problems of answer-sheet marking, etc., might go through four or five forms, each with experimental tryout on small numbers of cases, before item analysis was undertaken.

Item analysis was considered to be a very important tool. Experimental forms of a test were almost uniformly made long enough that considerable item selection might be done. Tests with high internal consistency were desired for factor-analysis purposes and for potential inclusion in the classification battery. It was realized that this was not necessarily the best way to maximize the validity of the individual test. Maximum validity for a single test was neither necessary nor desirable, however, since maximum validity of the battery of tests was the goal.

It should be emphasized that high internal consistency was desired for more than reliability alone. For one thing, items that have low correlations with the total score of which they are a part are not necessarily unreliable items. A low correlation with total score often indicates that the item measures some other function than that measured by the rest of the test. High internal consistency was desired because it increased the chances of obtaining a pure test. Items of low internal consistency with promise of validity posed a problem for additional test construction, that of finding a test in which they would belong.

Item analyses were used in ways other than for item selection. A considerable amount of item revision often occurred at this stage in the development of the test. The item analysis not only furnished the correlation between item and total score on the test, but it also furnished information concerning difficulty levels, functioning of misleads, and the extent to which the test was speeded.

Correlational and Factor Analysis

After item selection and revision had been accomplished, time limits revised, and directions given a final polishing, all the tests in the subarea were prepared for correlational administration. This administration often involved difficulties that could not always be overcome. No formal provision had been made for such testing. One or two experimental tests could be given along with the classification battery, but there was not sufficient time to give an experimental battery. The time of the aviation student, both during the classification process and during preflight training, was rather closely scheduled. The most desirable group would have been composed of unclassified students. Often, however, the only students available for extra experimental testing had already been classified. In certain factor analyses to be reported later in the volume, based on classified students, the results are somewhat biased as compared to those that would have been obtained if unclassified students had been used.

No matter what the source of subjects happened to be for a given analysis, classification-test scores were always available from the regular administration. A selection of the best known of these was made for inclusion in the matrix of correlations, to serve as reference tests. This procedure insured that certain known factors would be included in the analysis and would be readily identified.

It was at this stage, also, that reliabilities were usually computed. The principal use to which reliability estimates were put, as a matter of fact, was, in comparison with communalities, to obtain an indication of the amount of nonerror specific variance in a test.

Validation

On the basis of the data accumulated in the preceding stages, a ranking was made of the experimental tests with regard to their desirability for immediate validation. Promise of unique contribution was, of course, the chief criterion employed. Such a ranking was necessary because during most of the period covered by test research the amount of testing time allotted for experimental testing was limited. To obtain as many as a thousand unclassified aviation students on every test was impossible, and samples of this size were barely sufficient for pilot validation only. The number of tests validated was increased sharply during brief periods when preflight graduates were tested.

CONCLUSIONS

In this chapter a generalized picture was sketched of printed-test construction. It was seen that from the first, test construction was oriented toward the development of tests of functions or factors rather than toward job-sample-type tests. This stemmed from the analysis of Faculty Board proceedings which was couched in terms of traits of unsuccessful pilots. This analysis of the important traits necessary in learning to fly "the Army way" has been considerably modified and enlarged by subsequent factor-analysis findings. As validation findings and factor-analysis results became available, the direction of test research became progressively less influenced by job-analysis information.

The importance of constant and rapid validation of experimental tests was stressed. As a basis for evaluating the test validities to be presented in later chapters, the pilot, navigator, and bombardier criteria were briefly discussed. Reasons for the concentration of test research in the pilot area were also discussed. These are, in brief, as follows: The greater ease and promptness of validation against the pilot criterion; the importance of the pilot problem as a function of initial low validity in this area and large quotas; the initial high validity for tests against the navigator criterion; and the lack of reliability of the bombardier criterion.

The final section discussed the typical history of an aptitude test, proceeding from selection of the subarea, formation of test ideas, and item writing and criticism, through experimental administration, item analysis, and correlational and factorial analysis, to final validation.

CHAPTER THREE

Commonly Used Statistical Procedures[1]

Most of the steps in the typical history of a test discussed in the preceding chapter involve statistical computations of one kind or another. Report No. 3 of this series describes statistical techniques employed in all aspects of the AAF Aviation Psychology program, so it is unnecessary to go into detail here regarding those techniques. Certain techniques were selected as standard for use in the development of printed tests, however, and so it is desirable to set forth an account of the adaptation of those particular methods—to account for the choice of methods, to mention any special variation of them (for there were some), and to set down conclusions based upon extensive experiences with them. This chapter will also serve the purpose of explaining the nature of most tabular material in the chapters that follow, as well as the nontabular statistics used in describing tests.

RELIABILITY

Reliability has usually been defined as the correlation between comparable or interchangeable measures of the same thing. Other than to point out that the use of the singular word "thing" may legitimately cover a factorially complex test—that is, comparability does not imply item-for-item correspondence within a test, but merely from one form to the other—one does not need to amplify this definition in any way. Reliability as thus defined is a useful concept in test analysis. In most cases, also, the definition unequivocally suggests the appropriate technique of estimation.

Correlation between Comparable Forms

The technique of reliability estimation that has been most commonly used in printed-test development is a part I-part II correlation.[2] It involves computing the correlation between separately timed but comparable parts of a single test printed within a single booklet and administered in immediate succession. This procedure differs from the usual one involving comparable forms in two particulars: (1) Comparable forms are usually printed as separate booklets, and (2) are usually administered with a time interval between them. Some test technicians

[1] Written by Capt. Lloyd G. Humphreys.

[2] In the tables of this volume, this is referred to as an alternate-forms type of reliability.

believe that the intervening time interval is desirable, since it would presumably take into account function fluctuation within the individual from time to time, as well as function fluctuation during the course of the test. Data are available, however, on four rather different tests, which show that the reliability estimate is not significantly affected by the difference between immediate and somewhat delayed administration of the second part.

These four tests were administered in separately timed halves, and with two time-interval conditions. In the first condition, the second half was administered immediately after the first. In the second condition, about 4 hours of time and approximately half the tests in the group-test classification battery intervened between the 2 halves. The tests were selected as ones thought likely to show a decrement in reliability after an interval, if such a decrement does indeed occur. The tests chosen (*a*) were speeded and (*b*) called for a rather complex and novel task, involving extensive instructions. The tests were administered in pairs. A given group received one test of the pair without appreciable interval and the other with the 4-hour interval, and then the conditions were reversed for the next group. Approximately 1,000 cases were tested with each pair of tests, 500 in each sequence. The results are shown in table 3.1.

TABLE 3.1.—*Experimental test-reliabilities with and without time interval between parts I and II*

Test[1]	Statistic	Preaviation cadets only		Preaviation cadets plus airplane mechanics	
		Interval	No interval	Interval	No interval
Decoding, CI214AX2	N	238.	355.	426.	439.
	M_1	12.33	10.97	10.76	10.47
	M_2	14.20	13.22	12.31	12.48
	$S.D._1$	6.34	6.19	6.40	6.39
	$S.D._2$	6.60	6.64	6.79	6.85
	r_{12}	.58	.58	.64	.63
Estimation of Length, CP631A	N	355.	238.	439.	425.
	M_1	16.50	18.58	16.11	17.66
	M_2	11.61	12.50	11.62	11.97
	$S.D._1$	7.22	7.48	7.10	7.35
	$S.D._2$	6.64	6.51	6.50	6.46
	r_{12}	.41	.40	.46	.43
Object Identification, CP521A	N	524.	193.	......	448.
	M_1	46.94	48.13	......	41.02
	M_2	42.64	41.39	......	36.09
	$S.D._1$	14.74	16.19	......	18.12
	$S.D._2$	12.55	12.71	......	13.96
	r_{12}	.60	.65	......	.64
Visualization of Maneuvers, CI657CX1	N	193.	525.	448.	......
	M_1	16.12	20.36	11.72	......
	M_2	17.79	19.75	12.35	......
	$S.D._1$	10.02	10.74	10.15	......
	$S.D._2$	10.85	11.20	11.13	......
	r_{12}	.82	.85	.84	......

[1] For description of these tests see chapter 7, Reasoning tests; chapter 18, Size and Distance Estimation Tests; chapter 19, Spatial Tests; and chapter 12, Visualization Tests.

The intervening time interval and activities in this study are thus seen to have no measurable effect on reliability estimates. While it is possible that a longer delay, or other types of activities might produce such an effect, it should be noted that the delay and activities chosen represent the typical testing situation for correlational studies.

Advantages of the part I—part II technique.—Experience has shown that, for most tests, reasonably comparable forms or parts can be constructed without the use of elaborate trial forms and statistical analyses. Sophisticated inspection of the items placed in the two parts, if followed by a comparison of the two means and standard deviations, is usually a sufficiently rigorous technique. The labor involved in constructing two forms or parts is thus not excessive; printing in a single booklet reduces cost and inconvenience in administration; and having separately timed parts makes the method applicable to speed tests as well as power tests. As a matter of fact, this is the only satisfactory method applicable to both speed and power tests.

Odd-Even Estimates

In a few cases odd-even estimates of reliability were the only ones available, even on highly speeded tests. These are, of course, over-estimates of the reliability of speed tests. It is not so generally realized, however, that odd-even coefficients may underestimate the reliability of a power test, particularly if the test contains a small number of items, and if the test items measure different factors. When such reliabilities are presented, attention is called to their deficiency.

Use of the Spearman-Brown Formula

When the two parts correlated are truly comparable, i. e., when the product-moment correlation between paired items is 1.00 when corrected for attenuation, the Spearman-Brown correction gives a correct statement of the reliability of the two parts combined. The formula has been applied, however, in a number of cases where the two parts were not completely comparable. If the standard deviations of the two parts are not equal, application of the formula results in a slight underestimation of the reliability of the entire test. Lack of comparability of subject matter may result in grosser underestimates. Use of the Spearman-Brown formula will result in overestimates, on the other hand, when errors of measurement are correlated.

Uses for Reliability Estimates

Reliability estimation is not an end in itself. In a selection program one should be concerned about errors of measurement only as they affect validity. In a battery of tests it is usually more profitable to add a test of a new, valid function than to increase the length, and therefore the reliability, of one of the tests already in use.

It is useful, on the other hand, to have a reliability coefficient in analytical work with tests. Does the correlation between tests A and B represent all of their nonchance variances? How much would the validity of test A be increased if it were doubled in length? In any given factor analysis do a test's factor loadings account for all of its nonchance variance? Questions of this type can be answered knowing the correlation between comparable forms of a test. The greater the complexity the test, the more important it is in answering these questions to have item-for-item correspondence in the two forms, and the greater is the error involved in using any other estimate of reliability.

INTERNAL CONSISTENCY

Although reliability necessarily increases with increasing internal consistency of items, a reliable test is not necessarily internally consistent. It is possible to have a perfectly reliable test with zero correlations among its items, i. e., with zero internal consistency. This fact demonstrates the need for two concepts, and two terms in this area.

Kuder-Richardson Formulas

Of the Kuder-Richardson formulas (*8*) the one most widely used on tests in this volume is their No. 21, which involves the mean difficulty level of all of the items in the test. If the items do not vary widely in difficulty level, the error involved in not using the more accurate formula No. 20 is not great. With a wide range of item difficulties the latter formula is sufficiently precise. It makes the same assumptions as the analysis of variance, and in fact, is algebraically equal to Hoyt's formula (*5*) when the latter is applied to a test consisting of unit-weighted items.

Uses for Internal-Consistency Coefficients

Internal-consistency coefficients are often used as estimates of reliability coefficients. This must be done with care, however, since the two are only equal for a perfectly homogeneous test. Sophisticated inspection is an imperfect guide in using internal-consistency coefficients in this way.

The discrepancy between an internal-consistency coefficient and an estimate of reliability obtained from the correlation between comparable forms is somewhat indicative of the degree of heterogeneity of the test items. The larger the difference between the two, the greater is the degree of heterogeneity. This criterion is a sure indication of factorial complexity. The reverse is not true, however. If all items are factorially complex in themselves and to the same degree, the test will be both highly homogeneous and factorially complex.

Internal Consistency at the Item Level

The ultimate criterion of an item's consistency with the rest of the test of which it is a part is the level of its correlations with the other items. Since these correlations are unobtainable without excessive labor in most cases, some way of relating the item to total test score is used instead. Many methods of doing this have been suggested, but the approach to this problem has been characterized more by expediency than by rationality.

The phi coefficient.—The item statistic used on most of the tests in this volume is clearly in the expedient group, though it can be related more directly to a rational technique than most. The procedure used has been to compute the phi coefficient between passing or failing the item and belonging to criterion groups of the highest 27 percent and the lowest 27 percent of total score on the test. This procedure has a number of advantages. The group of papers is separated into high and low criterion groups at the outset; thereafter no further sorting of papers is necessary. With test responses recorded on standard IBM answer sheets, frequencies of responses to correct answers and misleads can easily be obtained by the use of the IBM scoring machine equipped with item-count attachment. The phi coefficients can then be read off a table or nomograph (*2*) after frequencies have been transformed into percentages or proportions. The statistical labor, compared with that in computing biserial coefficients, for example, is thus incomparably less.

The phi coefficient computed in this way has a number of interesting properties. The maximum phi is obtained at a difficulty level of 50 percent correct responses. For difficulty levels deviating from 50 percent, the maximum phi coefficients become progressively lower, while the sampling stability of the statistic remains unchanged. Since one's aim is usually to favor items near the 50 percent level of difficulty, the phi coefficient serves a double function in item selection. Use of this statistic alone therefore, tends automatically to produce a test of maximum internal consistency (*3*), and at the same time optimal difficulty. If, however, the test's specifications call for an appreciable number both of very easy and very difficult items, item difficulties will have to be considered as well as the phi coefficients.

The standard practice has been to compute two phi coefficients for every item in a test. One is based upon total groups; i. e., the computations are made on the assumption that omissions and items not attempted are wrong answers. This value is obviously in part a function of the item's position in the test, if speed is even a minor factor. It is therefore related to the internal consistency of the test as a whole under specified conditions of administration. The second is based upon total answered; i. e., computations are based on only those examinees who attempt the item. This coefficient is indicative of the internal consistency of the item only, independent, except for item interactions, of its position in the test. The distribution constants for item statistics presented in the chapters following are based on total answered in order to give as true a picture as possible of the items themselves. Another condition observed was that no phi coefficient was entered in these distributions, if it was based on less than 20 percent of the sample of cases in either criterion group. Thus, items near the end of a speeded test are not covered by these data. No attempt should therefore be made to relate the mean phi coefficients reported either to the standard deviation of the total-score distribution or to the internal-consistency coefficient for the test as a whole. These data are particularly inadequate for a highly speeded test, since one does not expect or desire individually discriminating items.

An empirical study of various item statistics.—Table 3.2 contains an empirical comparison of various item statistics. The purpose of the study was to compare the sampling stability of several commonly used item statistics in two representative samples of 400 cases each. The test analyzed was Visualization of Maneuvers, CI657CX1 (see ch. 12). Several measures of sampling stability are offered. The first of these is the correlation between comparable item statistics in the two samples. The second is a standard error of measurement computed as follows: $S.D.\sqrt{1-r_{I\,II}}$, in which SD is the standard deviation of the distribution of statistics over all items. The third is the critical ratio formed by dividing the mean item statistic by the standard error of measurement.

TABLE 3.2.—*Comparison of various item statistics on two samples of 400 classified pilots (68 items from Visualization of Maneuvers, C1657C were used in these analyses.)*

Statistic	Sample	M	SD	$r_{I\,II}$	SE[1]	[2]M/SE
Phi—27 percent	I	0.435	0.210	0.91	0.062	6.77
	II	.415	.203	....	....	....
Phi—50 percent	I	.298	.126	.87	.050	5.84
	II	.287	.150	....	....	....
Flanagan r	I	.463	.204	.87	.073	6.25
	II	.450	.202	....	....	....
Point biserial r	I	.358	.134	.88	.050	7.02
	II	.345	.158	....	....	....
Biserial r	I	.485	.168	.87	.065	7.24
	II	.456	.190	....	....	....
Tetrachoric r	I	.469	.180	79	.090	5.14
	II	.445	.212	....	....	....

[1] Computed from $r_{I\,II}$ and the average standard deviation in the two samples.

[2] The mean correlations entering into this ratio are, of course, spuriously high since an item is always correlated with a sum in which it is included. In view of the large number of items, the amount of error is very small, and is proportional to the size of the spuriously high mean correlations.

It is obvious from a comparison of the last three columns in table 3.2 that the question of sampling stability is answered somewhat differently by the three different criteria. If one were interested only in the rank order of item statistics in a second sample, there would be little basis for choice among the various item statistics with the possible exception of the tetrachoric correlation. If one were interested in a minimal standard error, a choice of either the point biserial or phi based on all the data would be clearly indicated. The writer is unaware, however, of any application where size of the standard error alone would be important. Lastly, if one were interested in detecting nonchance relationships, the two statistics that make use of all of the data would be the first choice, followed by those utilizing extreme criterion groups. These data, therefore, furnish empirical confirmation for Kelley's (7) theoretical formulation.

Since a biserial, continuous or point, is computationally laborious, a procedure using dichotomous-criterion groups is to be recommended. Discarding the middle 46 percent of scores on the continuous variable not only reduces the amount of item counting to be done, as compared to retaining all of the cases, but nonchance relationships are more easily detected as well. Choice of phi coefficients or Flanagan (1) r depends on whether one is interested in a statistic with a standard error independent of difficulty level or one in which the degree of relationship is independent of difficulty level. A statistic having the first of these two characteristics, i. e., the phi coefficient, has been found to have many advantages for item-analysis purposes.

The intercorrelations of the various item statistics for one sample only are presented in table 3.3. All statistics obviously are measuring much

TABLE 3.3.—*Intercorrelations of various item statistics for a sample of 400 classified pilots (68 items from Visualization of Maneuvers, CI657C, were used in computing these correlations)*

	Phi 50 percent	Flan. r	Point biserial r	Biserial r	Tetrachoric r
Phi—27 percent	0.91	0.98	0.92	0.88	0.88
Phi—50 percent	....	.90	.96	.88	.96
Flanagan r	....	....	.90	.90	.90
Point biserial r	....	....	....	.95	.94
Biserial r	....	....	....	....	.95

the same things. It is equally clear that more than one factor is involved. The two statistics computed on mutilated distributions are more like each other than they are like anything else, i. e., the correlation between the phi coefficient computed on upper and lower 27 percent groups and the Flanagan r is 0.98. The correlation between the tetrachoric correlation and the phi coefficient computed on upper and lower halves constitutes another doublet because they are computed from identical two-by-two contingency tables. The correlations with the point biserial are uniformly higher than those with any other measure, which indicates that the former may be the most representative item statistic in the group.

The point biserial correlation.—There are theoretical reasons why the point biserial correlation would be expected to be the most representative internal-consistency statistic. The point biserial, for example, can be most easily and directly related to the inter-item product-moment correlations, for which correlations of items with total score are substituted for reasons of computational convenience.[3] It can also be directly related to the standard deviation of the total score distribution, and therefore, to

[3] $\bar{r}_{ij} = \frac{n\bar{r}_{it}^2 - 1}{n - 1}$, where i and j are any two items, t is total test score, and n is the number of items in the test. $\bar{r}_{ij}$ is, therefore, the mean inter-item correlation, and $\bar{r}_{it}$ is the mean correlation between all items and total score. See appendix A for the derivation of this formula.

to the internal consistency of the test as a whole.[4] A simple expression of the amount of "bootstrapping" involved in the inclusion of item in total score can also be obtained through the use of the point biserial.[5] The highly satisfactory sampling stability of this statistic, shown by the three criteria in table 3.2, arises because all of the data are used and because it is a product-moment correlation, not an estimate of one, i. e., the percentages falling in the pass and fail categories do not affect its standard error. In fact the only consideration limiting the usefulness of this statistic in internal-consistency analyses is the computational labor involved.

Substitutes for the point biserial.—Use of the phi coefficient relating pass and fail on the item to upper and lower 27 percent criterion groups, as described in a preceding section, is a reasonably good substitute for the point biserial. The characteristics of the two are quite similar, and there is little loss in the efficiency with which nonchance relationships can be detected. Retaining all of the cases and computing the phi coefficient on a 50-50 split furnishes a statistic that is more highly correlated with the point biserial, but at a cost. The additional cost is the labor involved in obtaining item counts on almost twice the number of answer sheets, with a loss in the efficiency with which nonchance relationships can be detected.

The phi coefficient obtained from all of the cases grouped into upper- and lower-criterion groups has one advantage in that it is more flexible than either the point biserial or the phi coefficient obtained from the 27 percent criterion groups. The point biserial between item and total score is in part a function of the difficulty level of the test as a whole, in addition to being a function of the difficulty level of the item tested. The phi coefficient, analogously, is a function of the splits in both variables but the split in the criterion can easily be changed if no cases are discarded. If it is desired that the final test be relatively easy, even though the initial items given for try-out had an average difficulty level of 50 percent, the test-score distribution can be split at 75-25, for example. The maximum phi will accordingly be obtained, on the average, for items having difficulty levels of 75 percent, and the final selection of items will be biased in the desired direction.[6]

The procedure of "tailoring" a test for a particular cut-off point, i. e., selecting items of the same difficulty level as the percentage below the

---

[4] $r_{tt} = \frac{n}{n-1} - \frac{1}{(n-1)\,\bar{r}_{it}^2}$ where the various symbols have the same meaning as before, and r_{tt} is used for the internal consistency of the test as a whole. See appendix A for the derivation of this formula.

[5] $\bar{r}_{it} = \frac{1}{\sqrt{n}}$, when $\bar{r}_{ii} = .00$. That is, the mean correlation between item and total score including the item when the mean inter-item correlation is zero is equal to the reciprocal of the square root of the number of items. See appendix A for the derivation of this formula.

[6] It is possible that the use of extreme criterion groups might be adapted in the same way. The percentages to be used when some other difficulty level than 50 percent is desired have not been worked out.

cut-off, has not been widely used. There are a number of reasons for this. Most important, a short test can not very well be tailored for two different difficulty levels, one for the percentage of failures in the criterion, the other for the cut-off in selection and classification. Both of these difficulty levels, in addition, were unpredictable on a long-term basis. The criterion, eliminations in training, fluctuated widely without much regard for the ability level of the entering student. Although the cut-off point in selection and classification was raised progressively, the changes were relatively rapid and were usually made without advance notice. Most classification tests, in addition, are weighted for more than one specialty, each of which may require a different appropriate degree of test difficulty. Thus the advantage accruing when all of the cases are categorized, as opposed to using highest 27 percent and lowest 27 percent criterion groups, was not needed. Otherwise, the selection of the procedure involving omission of 46 percent of the cases from the middle of the distribution as the standard computational technique would not necessarily have been most advantageous.

Item Difficulty

In computing phi coefficients in internal-consistency item analyses, percentages of examinees passing an item in the upper and lower criterion groups are obtained. The average of these two percentages gives an estimate of the difficulty level of the item. For various reasons these difficulty levels can be considered merely approximate as long as item counts are made only in the tails of the total-score distribution. When based on a percentage of correct responses in a total group, the difficulty level of the item is in part a function of its position in the test. When based on a percentage of attempts, difficulty level is independent, except for possible item interactions, of position, but is biased by reason of selection of a sample of those who attempt many items in the test. People who attempt many items are usually those who are most able in the test as a whole. Values computed in the first of these two ways are used in Kuder-Richardson internal-consistency coefficients, since it is the internal consistency, and therefore the reliability, of the test as administered in which one is interested. Data concerning difficulty level of items presented in the following chapters on tests, however, follow the second procedure. Statistics based upon "total group" furnish more information about the test as a whole; statistics based upon "total answered" furnish more information about the items as such.

Correction for chance success.—In addition to being based upon total answered, the item-difficulty data in the following chapters are corrected in the conventional manner[1] for chance success. This procedure follows the usual reasoning to the effect that the expected proportion of chance

[1] $cp = \frac{np-1}{n-1}$, where p is the uncorrected proportion of passes and n is the number of alternatives (2).

success for a two-choice item is one-half, for a three-choice item one-third, etc., and that all examinees who do not know the correct answer guess at random. As a matter of fact, the amount of random guessing varies considerably from test to test, depending on the type of test and the care with which misleads are selected. Often, a given mislead is chosen by the examinee on the basis of misinformation, wrong hypothesis, or perceptual error. The greater the extent to which this is true, or in other words, the extent to which misleads and correct answer are not equally attractive to examinees who do not know the right answer, the larger is the amount of over-correction that results from the application of the formula based upon the random-guessing hypothesis.

Evidences against the guessing hypothesis—Various evidences are available concerning the inapplicability of the guessing hypothesis. These can only be briefly listed here. Difficulty levels for reliable items, for example, sometimes correct to zero or even to negative values. A test, secondly, can sometimes be made more internally consistent and reliable by an appropriate reduction in the number of misleads.[8] Lastly, the corrected difficulty level of an item or a test does not always remain constant, as the number and character of misleads is varied.[9] These considerations lead to the conclusion that item difficulty is not very closely associated with number of misleads in some tests. When difficulty values are given for tests in later chapters, therefore, there is a bias in the direction of overestimation of difficulty rather than underestimation.

VALIDITY

Validity data cited in this report are based on an extremely practical definition of the concept. The validity of a test is its relation to any variable one is interested in predicting. A test has potentially as many validities, therefore, as there are criteria available.

Validation Statistics

It was pointed out in the preceding chapter that the most common criteria of success as pilots, navigators, or bombardiers were pass-fail data. The prediction of graduation or elimination is the most useful datum to those concerned with training problems; information concerning graduation or elimination was also easiest to obtain of all criteria available.

The biserial correlation coefficient.—With test data for all practical purposes continuously distributed and a dichotomous criterion, the biserial correlation coefficient is immediately suggested. This has indeed

[8] The mean phi in a variable 2-, 3-, or 4-choice 100-item test, Geography, AS104, was raised 0.023 (from 0.307 to 0.330) based on "total group" and 0.038 (from 0.271 to 0.309) based on total answered, changes that were beyond the 5 percent and 1 percent levels of significance respectively, over the mean internal-consistency values in the otherwise identical 100-item 5-choice version, AS102. This was true even though the mean phi of the 212 eliminated misleads was −0.041 based on "total group" and −0.068 based on "total answered"; i. e., the average eliminated mislead had discriminated between high- and low-criterion groups in the expected direction.

[9] The mean difficulty level of Mechanical Information, CI905A is 0.48 when corrected for chance success. The comparable value for CI905BX1, an experimental two-choice version identical with the first form except for the omission of three misleads in every item, is 0.38. This difference is beyond the 1 percent level of significance.

been the standard validation statistic used in the test-construction program. Its chief advantage is its independence of split in the criterion so that test validities computed at one period of time can be compared with validities computed at a later date without regard to differences in elimination rates. The fact that the biserial correlation also gives an estimate of what the product-moment correlation would have been if the criterion had been continuously and normally distributed is perhaps satisfying, though whether this constitutes an advantage in prediction is debatable.

Use of the biserial correlation coefficient is subject to one serious drawback. Formulas commonly used to correct for restriction of range are not strictly applicable to biserial correlations. The greater the amount of the restriction and the greater the disparateness of the split on the dichotomy, the greater is the error involved. Because of the high number of disqualifications for low aptitude at the time of classification, navigator validities were somewhat underestimated almost from the start; pilot and bombardier validities were significantly underestimated only after many months of the test-construction program, as the number of low-aptitude disqualifications increased. For the degree of restriction of range due to the elimination of the lower 60 percent or more of scores on the various stanines[10] and for the elimination rates commonly encountered in flying training, the amount of error in the corrected biserial correlation may amount to as much as 0.10.

The point biserial.—An alternative statistic for use in the validation of a test against a dichotomous criterion is the point biserial. The applications of this statistic to psychological problems have not been sufficiently investigated to make definite recommendations. One disadvantage is immediately suggested—the point biserial is in part a function of the split in the dichotomous criterion. In order to be compared, test validities would have to be equated for differences in elimination rates. The fact, however, that the point biserial is a product-moment correlation, not an estimate of one, suggests that it might be useful.

Validation Procedures

Experimental tests were most often given to unclassified aviation students. Ideally, the aviation students should not be able to distinguish an experimental test from a classification test on which their qualification and classification depended. Thus, much validation-test administration was conducted along with classification testing. The first step in the validation procedure was, therefore, to obtain the classification records. If one were interested in pilot validation, the classified pilots were separated from navigators, bombardiers, etc. After the necessary interval of time, rosters of graduates and eliminees from elementary pilot training were searched[11] for men to whom the test was administered.

---

[10] A stanine is a standard score, on a 9-step scale with a mean of 5.00 and a standard deviation of 2.00, which represents the composite aptitude score for a given type of flying training.

[11] Readers who are familiar with punch-card techniques can immediately translate this and other steps in the procedure to jobs of sorting, collating, tabulating, etc.

Selection of elimince group.—Of the several types of eliminees listed by training schools, e. g., flying deficiency, academic deficiency, fear of flying, traits of character, and physical deficiency, nonpsychological categories constituting a small percentage of total eliminees were consistently omitted in the computation of the biserial correlations. In periods of high elimination rates, in addition, pilot validation was often accomplished on the basis of flying-deficiency eliminees only. This was done because flying deficiency was presumably the purest criterion available of pilot failure. Data are available, however, which show that stated type of elimination from pilot training is relatively unimportant as measured by the degree of relationship between the composite aptitude score (stanine) and the various categories of eliminations.

The correlations obtained.—In addition to computing the biserial correlation between the test and pass-fail in training, certain other correlations were routinely obtained as well. The biserial correlation between the appropriate stanine and the criterion was obtained for the same sample as a partial check on the representativeness of the sample. A product-moment correlation was obtained between the test and stanine in order to estimate the amount the test might add to the predictive efficiency of the classification battery. The latter two correlations are not presented routinely in this volume because a stanine does not constitute a fixed reference point, since the composition of stanines was changed with each new classification battery. Both correlations are used, however, in correcting a test validity for restriction of range. These corrected values are routinely presented for tests administered after the time that restriction of range became a serious problem.

Correction for restriction of range.—The standard formula[12] for correcting a test validity for restriction in range is derived from the formula for the partial correlation coefficient. Restriction in range of a variable (the stanine) has, as a limit, the reduction of the variable to a constant, with effects on related variables (test and criterion) predictable from the partial correlation technique. In a real sense, therefore, the partial correlation coefficient is a special case of restriction of range.

The effect of restriction of range on a test's validity depends upon its correlation with the stanine. It is not difficult to see why a test's validity should be increased when correcting for restriction of range when the test is highly correlated with the stanine. The correction will often change a small negative validity to a positive one if the correlation with stanine is substantial. Unless the relationship of the correction formula to the partial correlation coefficient is remembered, however, it may be

---

[12]

$$r_{12}'=\frac{r_{12}+r_{13}\,r_{23}\left(\frac{S_3^2}{s_3^2}-1\right)}{\sqrt{\left[\left(1+r_{13}^2\frac{S_3^2}{s_3^2}-1\right)\right]\left[1+r_{23}^2\left(\frac{S_3^2}{s_3^2}-1\right)\right]}}$$

where S_3 is the unrestricted standard deviation of variable 3; s_3 is the restricted standard deviation on the sample, and the correlation corrected for restriction in range of variable 3 is r'_{12}. See report No. 3 of this series for a more complete evaluation and description of corrections for restriction of range.

surprising to find a test's validity lowered by applying the correction formula when the test is relatively uncorrelated with the stanine. A number of instances of this sort will be found in the chapters to follow.

Most of the corrected validities reported in this volume are based upon an assumed standard deviation of the unrestricted stanine of 2.00. This follows, of course, from the definition of stanine. During a period of several months, however, it was believed that in the process of setting up conversion tables of raw aggregate scores to stanine units, standard deviations of stanine significantly less than 2.00 had been obtained. Certain validities are reported as corrected to values of the unrestricted stanine of less than 2.00. The actual value used in making the correction is consistently footnoted.

A further source of confusion in making corrections for restriction of range is that at times data were available for the augmented pilot stanine only. The pilot stanine was augmented by adding either two or three stanine points to the aptitude score of students with specified amounts of previous flying experience.[13] The proportion affected by this procedure varied from time to time, but was usually in the range from 0.10 to 0.15. The standard deviation of the augmented stanine was therefore greater than that for the unaugmented, the difference usually being about 0.10. The assumed standard deviation of the unrestricted augmented stanine was accordingly 2.10 during most of the period covered by this volume. The exceptions noted above for the unaugmented stanine, however, also apply to the augmented.

Item Validation Procedures

Item validation was used primarily in the selection of items and for keying responses in personality inventories, biographical data blanks, etc. Ease of computation is always a criterion in any work with items, but beyond that is the need for a statistic that will produce valid empirical keys.

The tetrachoric correlation coefficient.—Since item validation necessarily involves a two-by-two contingency table when the criterion is dichotomous, the tetrachoric correlation is suggested as a suitable statistic. It was used in fact in a number of item-validation studies. Its chief advantage is that the degree of correlation is independent of varying elimination rates or item difficulties. Its standard error, on the other hand, is a function of both these variables. For sampling reasons, high correlations tend to occur predominantly on very easy or very difficult items. Unless some criterion of statistical significance is used in addition to the correlation, item selection will be biased away from items of moderate difficulty level towards very easy or very difficult items. Such items are less likely, because of sampling errors, to give comparable results on a second administration and do not afford maximum discrimination among

[13] A maximum stanine of 9 was adhered to in spite of additions, however.

703330—47—4

examinees. For these reasons the tetrachoric correlation was not used as widely as the statistic to be described.

The phi coefficient.—The phi coefficient can be computed from the same contingency table as the tetrachoric correlation. It has quite different properties, however, since its standard error is independent of split in either dichotomy, while the size of the correlation is a function of split. Assuming a constant level of item intercorrelations, the mean phi coefficient between test items and the criterion can be directly related to the point biserial correlation between total test score and the criterion.[14]

For an item that discriminates positively, phi is at a maximum when the number marking a given alternative equals the number in the superior criterion group (graduates). For maximum negative discrimination, however, phi is at a maximum when the number marking a given alternative equals the number in the inferior criterion group (eliminees). If this statistic were used unmodified, items selected for keying at one level of the graduation rate would not be the best items to use if this rate were to change radically.

The computation of phi coefficients was slightly modified in practice as a means of partially overcoming the dependence of the statistic on a given graduation rate. All item-validation statistics were computed on the assumption of equal weighting of graduate and eliminee groups; i. e., the same charts were used for item validation as for internal-consistency computations, where the upper and lower criterion groups were always of equal size. This procedure has the effect of increasing all item phi coefficients, but increasing most those for splits furthest removed from the one where phi would be at a maximum in the more precise procedure. It follows, therefore, that the application of a standard-error formula is an approximate procedure, although a single standard error can be used more precisely for phi computed in this way than for the tetrachoric correlation.

No perfectly satisfactory item validation statistic was used. If the graduation rate from training were more constant, the phi coefficient computed from the actual split in the criterion would be less open to criticism. With changing graduation rates the "compromise" procedure may be more generally useful.

Cross validation.—The development of a key for a personality or biographical-data inventory can be very briefly summarized. The first step involves the experimental administration of the inventory. This was done either during the classification period or while students were in preflight training, or at the time of graduation from preflight training. The accumulation of 2,000 cases of classified students would have been desir-

[14] $$r_{tc} = \frac{\bar{r}_{ic}}{\bar{r}_{it}},$$ where the subscript *c* refers to the criterion, and the other symbols have the same reference as before. That is, the correlation between total score on a test and the criterion is a function of the ratio of the mean item validity to the mean internal-consistency coefficient.

able, although this was not always possible. When graduation-elimination information became available, the unscored answer sheets were separated into those for graduates and those for eliminees. The answer sheets of these graduate and eliminee groups were then divided into odds and evens, usually on the basis of odd and even testing numbers. Item counts were obtained, as a next step, and item validities were computed for odds and evens separately. Using responses exhibiting a difference at or beyond approximately the 5 percent level of significance, and avoiding the alternatives selected by a very small (usually less than 0.15) or very large proportion of the cases, separate scoring keys were made up for the two samples. The odds answer sheets were then scored with the evens key and vice-versa. Validation statistics were then computed in the way already outlined.

The cross-validation procedure avoids the "bootstrapping" involved in scoring answer sheets on a key prepared from item-validation statistics computed on the same sample. If both the odds and evens keys are valid, the final recommended key obtained from combining the two experimental keys will, on the average, be more valid.

FACTOR ANALYSIS

The previous chapter discussed the importance attached to factor analysis in test construction research. In the present section will be discussed some of the statistical and computational aspects of the technique. In this connection several factor-analysis schools will be briefly covered.

Common Assumptions

Factor analysts, no matter how much they may differ among themselves on certain points, make common assumptions in their factor solutions. Any given distribution of test scores, for example, is assumed to result from a weighted additive combination of orthogonal[18] reference factors. The correlation between any two variables, therefore, is also an additive combination. It can be written as follows: $r_{12}=a_1a_2+b_1b_2+\ldots k_1k_2$, where a_1 is the loading in factor a of test 1, etc., and where k is the last factor in the analysis.

Many critics of factor analysis have seized upon the additive assumption as a possible weak link. The additive assumption does not allow for complex interactions of parts, for the whole being unpredictable from knowledge of the parts, or for parts being unrecognizable in the whole. This is a question, however, to which an experimental answer is possible. Evidence is presented in chapter 28 showing to what extent the additive assumption has been found to correspond to test data.

Divergent Computational Procedures

Any casual statistical reader knows the centroid method of factor analysis by name and associates it with Thurstone (9). Almost equally

[18] Thurstone departs from orthogonality, for example, only in the rotational process.

well known are the principal-axes method of Hotelling (4), and the principal-components method of Kelley (6). This same reader knows that there are certain disagreements among these individuals concerning methodology, but is usually quite uncertain concerning the actual points of disagreement.

The mathematical solutions.—There is little difference of opinion concerning the mathematical solutions. Either the principal-axes method or the principal-components method is superior to the centroid solution mathematically, but inferior from the standpoint of computational labor involved. The mathematical superiority of the first two is due primarily to the fact that each succeeding factor extracts a maximum portion of remaining variance.

It is often stated that the first two methods are not scientifically parsimonious; i. e., because, as commonly used, as many factors are extracted as there are variables in the correlational matrix. This criticism is not justified. If all methods are applied to the same correlational matrix with the same diagonal entries, there is no difference in parsimony in favor of the centroid method. Use of 1.00 in the diagonal results in as many factors (not necessarily all reliable) as variables. If the communality is used as the diagonal entry instead, no more factors need be retained by one method than by any other. If the factors computed by various methods are compared by ordinal number, the centroid factors will be found to reduce the variance in the matrix less sharply than the other two methods.

Diagonal entries.—The real crux of the differences among factor analysts lies in the selection of the diagonal entry in the correlational matrix. This in turn is directly related to the problem of whether to rotate axes.

The advocates of the use of 1.00 as the diagonal entry seem to value most highly the mathematical advantages that accrue when this procedure is used. Being able to assess the reliability of a factor is indeed a considerable advantage. The writer, among others, is unable, however, to find many psychological advantages in this procedure.

If 1.00 is used as the diagonal entry in a correlation matrix composed of coefficients uncorrected for attenuation, the resulting factors and factor loadings constitute an inextricable mixture of common factors, nonerror specifics, and error specifics. These factors are probably not very meaningful, although they furnish an exact mathematical description of the original correlations. Rotations are attempted only rarely on such data. It is the writer's guess, however, that stable positions of axes cannot be found in these analyses.

The advocates of the use of the test's communality, i. e., the sum of the squares of the common-factor loadings, in the diagonal forego mathematical nicety for greater psychological meaning. Rotations can be made to positions of the factors that will reoccur in subsequent analyses, relatively independently of the constitution of the particular matrices. Experience has shown that these factors do have psychological meaning. For example, interpretation of stable factors will enable the test constructor, in revising tests, to increase or decrease a loading in a given factor at will. Although few data are available, it is probable that psychological interpretation of a factor can be related to job-analysis information sufficiently accurately to weight tests for a criterion in the absence of validation data for these tests in connection with that criterion.

Procedures Used in Army Air Forces Test Research

Because most matrices to be analyzed were relatively large, the centroid procedure was used for computational convenience. In order to obtain meaningful common factors, estimates of the communality were used as the diagonal entries. Communalities were estimated by selecting the highest coefficient in a column.

Centroid computations.—Centroid loadings were computed in the customary manner with one exception. The criterion used in reflecting signs was the algebraic sum of a column disregarding the diagonal rather than the mere number of negative signs in the column. The newer procedure insures positive sums and undoubtedly comes closer to maximizing table totals than the earlier one.

Every user of the centroid method finds himself perplexed by the problem of the number of factors to extract. No single criterion is sufficient. Most of the criteria that have been suggested do not allow the extraction of a sufficient number of factors to obtain stability of factor patterns in rotations. An objective criterion that has been found to be useful is the comparison of the standard error of a zero correlation with the product of the two highest centroid loadings. Factoring should not be stopped until the latter is at least as small as the former. The criterion that was actually given final weight, however, was quite subjective. Interpretability of the results is the only possible basis at present for choosing between, for example, 9 or 10 factors. In most cases the objective differences between two successive centroid factors are too slight, and the change too smooth, to make a confident decision concerning the exact number of factors.

Rotations.—Axes were always rotated in pairs. This was accomplished in various ways. At first the factors were plotted in order to estimate the angle of rotation, and rotated loadings were obtained by calculating machine using the trigonometric functions of the angle. With more experience the angle of rotation was estimated from the numerical values alone, and the rotated loadings were obtained as before.

The original procedure was time consuming; the second procedure involved a difficult visualizing process. With both, computers had difficulty with signs. As a result a new procedure was devised[16] that minimized most of the difficulties encountered with the previous ones. A pair of factors is plotted by projection, utilizing T-square and triangle, from the

[16] Chief contributor: S/Sgt. Wayne S. Zimmerman (10).

original plot of each onto a third sheet of paper. Rotations are made directly on the plot, and the new axes are used in succeeding rotations. The entire process is geometric. Numerical values of the various factor loadings are not involved from the time the first plots are made until the final rotated loadings are read off the sheets. Rotations are made more accurately and more rapidly than by any other method tried.

All the rotational solutions presented in this volume are orthogonal. Nonorthogonal solutions were not attempted. It seemed, in the first place, that those who use nonorthogonal solutions place too much confidence in essentially negative results. How can one be sure that the next group of tests will not change the correlation between two factors from +0.15 to 0.00? As a matter of fact, most of the intercorrelations reported among factors are so low that it hardly seems worth while to depart from orthogonality, even if one could be sure that the correlations were the true ones.

CONCLUSIONS

This chapter discusses many statistical procedures common to the test analyses and descriptions to be presented in succeeding chapters. On the basis of rather extensive experience with certain techniques, evaluations, and recommendations were also made in several instances. The topics discussed were: Reliability; internal consistency both of the test as a whole and of the items composing it; validity, again both of the test and its items; and factor analysis. A more complete discussion of research techniques will be found in Report No. 3 of this series.

TABULAR SYMBOLS

For the convenience of the reader, a list of tabular symbols commonly used in this volume is appended below:

N_t = Total number of cases in a sample.

p_g = Proportion of total sample graduated from the indicated phase of training.

M_g = Mean score of graduates.

M_e = Mean score of eliminees.

SD_t = Standard deviation of score distribution for the complete sample, including graduates and eliminees.

r_{bis} = Biserial correlation coefficient between test scores and the criterion, uncorrected for restriction in range on the stanine.

$_c r_{bis}$ = r_{bis} corrected for restriction of range on the stanine.

r'_{11} = Product-moment correlation between scores on separate comparable halves, separate comparable forms, or odd and even groups of questions, of a test.

r_{11} = r'_{11} corrected for length by the Spearman-Brown formula.

$M\phi$ = Mean phi coefficient.

$SD\phi$ = Standard deviation of the distribution of phi coefficients.

p = Uncorrected proportion of individuals passing an item.

$_cp$ = p corrected for chance success.

The reader will be able to interpret minor variants of these symbols as they occur in this volume.

One liberty is taken in tables. For convenience, product-moment correlations are frequently entered in a column for r_{bis}. When this occurs, appropriate footnotes are always made.

BIBLIOGRAPHY

(1) J. C. Flanagan, General considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution, *J. Ed. Psychology*, 1939, **30**, 674–680.

(2) J. P. Guilford, *Fundamental Statistics in Psychology and Education*, 1942.

(3) H. Gulliksen, The relation of item difficulty and inter-item correlation to test variance and reliability, *Psychometrika*, 1945, pp. 79–92.

(4) H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychology*, 1933, pp. 417–441; 498–520.

(5) C. J. Hoyt, Test reliability, estimated by the analysis of variance, *Psychometrika*, 1940, pp. 305–310.

(6) T. L. Kelley, *Essential Traits of Mental Life*, 1935.

(7) T. L. Kelley, The selection of upper and lower groups for the validation of test items, *J. Ed. Psychology*, 1939, **30**, 17–24.

(8) G. F. Kuder and M. W. Richardson, The theory of the estimation of test reliability, *Psychometrika*, 1937, pp. 151–160.

(9) L. L. Thurstone, *The Vectors of Mind*, 1935.

(10) W. S. Zimmerman, A simple graphical method for orthogonal rotation of axes, *Psychometrika*, 1946, **11**, 51–56.

CHAPTER FOUR

Tests of Intellect and Information[1]

THREE GENERAL TEST CATEGORIES

The presentation of research with printed tests is divided into three sections: intellect and information, perception, and temperament. This follows the original division of research responsibilities among Psychological Research Unit No. 3; Psychological Section, Headquarters, AAF Training Command; and Psychological Research Unit No. 1, respectively.[2] The coding system used for printed tests parallels this division of research responsibilities, and, therefore, the organization of the volume.[3] While the relationship between coding system and volume organization is not perfect, most of the tests in the CI series will be found in chapters 5 through 14, those in the CP series in chapters 16 through 21, and those in the CE series in chapters 23 through 27. The existence of exceptions to this relationship is indicative of the absence of sharp lines of demarcation between the categories, e. g., certain tests that were once given a perceptual or a temperament code number were considered to be more similar to intellectual tests when the volume was being organized.

Although other ways of categorizing tests might have been devised, the system used has the advantage of meaning much the same thing to most psychologists. In spite of some degree of overlapping of categories, there will be relatively few disagreements about the placement of most tests, although psychologists of different backgrounds may describe the categories quite differently. The following statements, in brief, constitute one such description. The intellectual category can be distinguished from the perceptual by the use of symbolization in the statement of the questions and misleads. Intellectual tests require symbolic mediation, usually by verbal or numerical symbols; perceptual tests do not. Both differ from temperament tests in that the latter stress manner or way of behaving, while the other two stress amount of knowledge or ability. The correct answer to an item in an intellectual or perceptual test follows inflexibly from a set of rules, while there is no right answer in this sense to an item in a temperament test.

[1] Written by Capt. Lloyd G. Humphreys.

[2] This division of research responsibility was later modified in a number of ways and was never adhered to perfectly. The code number of a test is, therefore, not a sufficient guide to the originating unit.

[3] The coding system was discussed in some detail in chapter 2. It will be remembered that the letter "C," which denotes a classification test, is followed by a second letter which indicates one of four main groups of tests: I—Intellectual, P—Perceptual, E—Temperamental, and M—Psychomotor.

HISTORICAL REASONS FOR THE USE OF TESTS OF INTELLECT AND INFORMATION

Academic Requirements for Flying Training

Beginning in 1920, high-school graduation or its equivalent was made a requirement for entrance into Army aviation training. Aviation flying training, at that time, consisted of training pilots only, although pilots studied navigation as well, frequently subsequent to graduation from flying school. In order to determine the qualifications of candidates who had not graduated from high school, administration of examinations was authorized. These examinations were expected to cover pertinent high-school subjects. By 1925, two such examinations were scheduled each year.

In 1927, the educational requirement was increased to 2 years of college or its equivalent, owing to the increase in the number of applicants for aviation training. Candidates who did not have 2 years of college training were given a special essay-type examination covering nine college subjects. This procedure was followed until the imminence of our involvement in war demanded a more extensive selection and classification program.

The Substitution of Objective Examinations

The need for a more objective and standardized qualification instrument led to a request by the Air Corps that the Personnel Procedures section of the Adjutant General's office construct an objective-type educational examination to be used for air-crew selection. This examination, completed late in 1941, consisted of five required sections, four of which were mathematical and the fifth, English composition. Five additional subjects—general history, United States history, physics, chemistry, and a language—were listed from which two options could be chosen.

After being in use for only 2 months, the objective educational examination was supplanted in January 1942 by the Aviation Cadet Qualifying Examination. A month later the testing program was extended to include a battery of classification tests which were administered to all who had qualified for air-crew training.

A natural consequence of the substitution of objective testing, in place of the prerequisite of 2 years of college, was the suggestion that tests of intellect and information be included in the classification battery. It was not clear at the outset whether 2 years in college or its equivalent was predictive of success as a pilot, bombardier, or navigator, though the case for the latter at least was fairly clear-cut. Studies of these relationships, however, constituted an obvious first step.

JOB ANALYSIS FINDINGS IN RELATION TO TESTS OF INTELLECT AND INFORMATION

The Pilot

The analysis of faculty-board proceedings, mentioned in chapter 1, resulted in the establishment of four categories in the area of intellect and information. These categories were generally defined in terms of expressions used by instructors in describing reasons for elimination.

Judgment.—Ability to make sound judgments and choices as to the best thing to do when faced with practical problems in traffic, in making forced landings, and in similar situations.

Foresight and planning.—Ability to plan a series or sequence of maneuvers, plan ahead for landings, plan entry into or exit from traffic, and foresee and avoid possible difficulties.

Memory.—Ability to remember instructions from day to day, both general explanations and specific, detailed information.

Comprehension.—Ability to grasp the meaning of explanations, instructions, and demonstrations when they are given either orally or in written form.

In spite of the fact that comments appeared in one or more of these categories in reports of 68% of all eliminations from pilot training, it was realized at the outset that the importance and psychological uniqueness of these categories remained to be established. Relatively little dependence can be placed upon descriptions of psychological traits made by untrained personnel, particularly when a limited and informally standardized vocabulary is used to characterize failures.

In addition to the categories yielded by the analysis of faculty-board proceedings, early job-analysis information indicated rather strongly the importance of mechanical information and mechanical comprehension for the pilot. The mere fact that the airplane, of which the pilot has charge, is an extremely complicated mechanism was sufficient basis for starting research to determine the pilot validity of mechanical tests.

The Navigator

In navigator training, the one thing that stood out in even the most cursory job descriptions was the importance of numerical and mathematical skills. In general, the task of the navigator seemed to call for the same traits that are necessary for success in academic pursuits. While this conclusion has been somewhat modified by subsequent experience, it is still true that the navigator is the most academic member of the air crew. Certainly, tests of intellect and information were high on the priority list for research on the problems of navigator qualifications.

The Bombardier

Early descriptions of the job of the bombardier were so meager and so conflicting that relatively little basis was furnished for test construction. There was some consensus that mechanical tests might be valuable in the selection of candidates for bombardier training. Since later studies showed the bombardier criterion to have little reliability, conflicting reports concerning the qualities of good and poor bombardiers were to be expected.

THE CODE NUMBER SYSTEM

Based on the available job-analysis information, a code-number system was established for tests in this area. Successive hundreds in the CI series are assigned as follows:

100 Information.
200 Reasoning.
300 Judgment.
400 Foresight and Planning.
500 Memory.
600 Comprehension.
700 Mathematics.
800 Physics.
900 Mechanical Comprehension.

Mathematics and physics are clearly not coordinate with the other categories. They were listed separately because of the expected volume of tests under those headings. Since later test construction was closely geared to validation studies, the expected volume of tests in certain areas, e. g., physics, did not materialize. There are, also, relatively few information tests as such. Information tests have been designed largely as interest tests, and so their assigned code numbers have been in the temperament (CE) area.

CHAPTER ORGANIZATION

The organization of the chapters in this section follows the coding system with relatively few exceptions. The closest correspondence between chapter content and coding system is for the following: Chapter 5, Verbal Ability Tests, all in the comprehension area; chapter 6, Mathematics Tests; chapter 7, Reasoning Tests; chapter 9, Foresight and Planning Tests; and chapter 11, Memory Tests.

Other chapters correspond very closely to the logical framework of the coding system, but exhibit minor irregularities. Chapter 8, Judgment Tests, includes, besides judgment tests *per se*, tests of estimation and of fluency. Construction of these other types of tests was based on hypotheses concerning the unique components in the factorially complex act called judgment. Chapter 13, Mechanical Tests, includes a discussion of physics tests. The physics tests were too few and too lacking in importance in the test construction program to warrant a separate chapter. There is, in addition, an obvious relationship of physics tests to the mechanical area.

Chapter 14, Information Tests, is irregular in that it includes tests that, as presumptive measures of interests, were given temperament code numbers. The decision to include these tests in this section rather than later was somewhat arbitrary. To have done otherwise would have divided information tests between two chapters. The factorial content of these tests is prevailingly intellectual rather than temperamental, in spite of the test constructors' intentions. The provision for an information category in the CI area is another argument for including the information tests in this section.

Chapter 10, Integration Tests, constitutes one of the two major exceptions to the relationship between coding system and chapter organization. The integration concept arose from two different sources. The ability to integrate the influence of several simultaneously operative elements in a situation, all of which bear upon the choice of a single direction of action, seemed an apt description for a valid pilot factor discovered in one of the early factor analyses. Later job descriptions also furnished additional evidence of the importance in pilot training of this hypothesized ability. Of the available categories, comprehension seemed most like this concept. The integration tests were therefore given code numbers in the latter half of the comprehension series.

The second exception is Chapter 12, Visualization Tests. This chapter is close to the borderline between this section and the one on perception. This is readily apparent from the variety of code numbers included in the chapter. Historically, the first visualization test in the program was constructed as one of a battery of mechanical-comprehension tests. The visualization ability, as a matter of fact, seems to be an important component of many seemingly mechanical tests. In a later battery of reasoning tests, the visualization factor again was found to be prominent. In this battery it was also discovered that a good visualization test could be presented entirely in verbal terms. Psychologically considered, visualizing is symbolic activity rather than direct response to sensory stimulation. These various evidences seemed to provide sufficient justification for including the visualization chapter in the section dealing with tests of intellect and information.

CONCLUDING STATEMENT

The informed reader will recognize many standard tests in the chapters to follow. Assignment of credit for construction of these tests is based upon the work necessary to adapt the tests to an aviation-student population or to an IBM answer sheet. The latter task, particularly, often requires no small degree of ingenuity. Because of the great amount of "tilling" the field of intellectual and information tests has had since the first Binet-Simon test scale was published, it has been difficult to make truly original contributions in test construction *per se*. The original types of tests, and there are several, are all the more gratifying for this reason.

These chapters contain, in general, more contributions to our knowledge about tests and the abilities they measure than they do descriptions of original tests. It is believed that the reader will be impressed, as the writer has been, with the need for a redefinition and reanalysis of general intelligence and a reconstitution of the instruments that purport to measure it. Most of the variance of standard tests of intelligence could undoubtedly be accounted for by appropriately selected and weighted tests taken from the three chapters on Verbal Abilities, chapter 5; Mathematics, Chapter 6; and Reasoning, chapter 7. Yet, such a combination would represent at least three relatively independent abilities in the aviation-student population, and any given intelligence-test score would therefore represent only one of several possible combinations of ability levels in the three traits. If differential validities of these traits exist for various job criteria, it is obvious that needless errors in prediction are made by using tests of general intelligence. And, although the user of such tests may thoroughly understand that no claims can be made for a complete coverage of human traits, he is likely to neglect many important abilities because of the social importance at present attached to intelligence. A more analytical approach would make such neglect virtually impossible.

CHAPTER FIVE

Verbal Ability Tests[1]

INTRODUCTION

Army Aviation Selection in Prewar Years

The development of tests of verbal ability was a natural outgrowth of studies carried on before the emergence of the aviation-classification program. As was stated in chapter 4, beginning in 1927, 2 years of college training, or the equivalent, were required for acceptance of applicants for flying training. Those who had not satisfied the college requirement could qualify by passing a special examination on nine college subjects. This early emphasis on scholastic achievement as a criterion for selection proved to be a forerunner of the Reading Comprehension and Vocabulary tests which were later employed in air-crew classification.

The AAF Qualifying Examination

The first two parts of the initial form of the Qualifying Examination are Vocabulary and Reading Comprehension. The following reasons were given for the inclusion of the vocabulary section:

> The purpose of the vocabulary section is to make possible the selection of men who have good general intelligence and are able to comprehend and understand written directions. Vocabulary tests have been found to predict the ability to understand and remember the sort of material that is covered in air-crew ground schools, where the student must remember what he reads and hears (2).

The following reasons were given for the inclusion of the reading comprehension section:

> The purpose of this section is to select individuals who can read and comprehend the sort of material that they must study and apply in aviation training. This section, like the vocabulary section, is a measure of general and intellectual ability (2).

Statistical results.—Statistical results soon revealed that the different parts of the test were of varying importance for predicting success in pilot, bombardier, and navigator training. The vocabulary section was of special value only for the prediction of success in navigation school (see table 5.1).

The reading-comprehension section showed a positive correlation with success in pilot training and a very high correlation with success in navigation training. In addition, it was the most effective part of the examination for predicting success in bombardier training (see table 5.1).

[1] Written by T/Sgt. Sanford J. Mock.

TABLE 5.1.—*Validation data for Vocabulary and Reading Comprehension tests for air-crew training*

Group	Test	N_t	p_s	M_s	M_e	SD_t	r_{bis}
Pilots[1]	Vocabulary	545	0.59	28.80	28.61	7.16	−0.04
Pilots[1]	Reading comprehension	545	.59	12.49	12.03	2.02	.14
Navigators[2]	Vocabulary	221	.79	31.87	27.85	7.03	.32
Navigators[2]	Reading comprehension	221	.79	13.05	11.13	2.15	.52
Bombardiers	Vocabulary	191	([3])c	29.56	27.39	6.09	.18
Bombardiers	Reading comprehension	191	([3])c	12.38	11.06	2.20	.31

[1] In class 42-G. Tested at Psychological Research Unit No. 1, January 1942.
[2] In classes 42-6, 42-7, and 42-8. Tested at Psychological Research Unit No. 1, January 1942.
[3] Not reported.

Verbal-Ability Requirements of Air Crew

Knowing the complex nature of the airplane and its operations, we may reasonably suppose that greater-than-average intellect would be required for success as an air-crew officer. All air-crew members become officers on attaining their wings. Some attention, therefore, had to be paid to the selection of potential commanders—men possessing superior leadership ability and intellect. It was logical, therefore, to seek a known measure of so-called general intelligence. This led at once to tests of verbal ability and comprehension, for, as Bingham (1) points out:

> Without recourse to language, the processes of comparison, abstraction, generalization, and mental organization would be limited indeed. With the aid of verbal symbols, we can more easily wrestle with problems, manipulate meanings, and test possible solutions of our difficulties mentally before we act. Little wonder, then, that a good test of vocabulary is of use as an indirect measure of a person's verbal or conceptual intelligence, and for two reasons: First, the richer his store of words and meanings, the better his equipment for solving some of his problems promptly and correctly, that is, for showing intelligence; second, the more intelligent he has been since infancy, the greater the likelihood that he has gained command of a wide variety of correct word meanings. Intelligence is far from being identical with the power to read understandingly, to speak aptly, or to write coherently and concisely. But the reciprocal relations between mastery of the mother tongue and ability to think intelligently should be obvious.

Need for language proficiency in ground school.—Although a poor showing in ground-school courses might not be sufficient basis in itself for elimination from primary pilot training, nevertheless, failure to grasp the theoretical concepts of flight would surely limit an individual's understanding of the function of an airplane and would possibly affect his performance in the air deleteriously. Bingham (1) points out a common cause of failure in school subjects.

> Mention has been made of yet another danger signal: Poor ability in English. A lack of equipment in the verbal tools of thought, revealed by low scores in tests of vocabulary and of English usage, may signify either insufficient training in the clear and precise use of language, or a shortage of verbal intelligence without which it is difficult to master college subjects. * * * The candidate's previous school achievements and his performance in scholastic aptitude tests furnish evidence regarding his general mental ability.

The navigation course includes much theoretical technical material that must be read and comprehended, in such subjects as calibration, radio navigation, and celestial navigation. Intelligence is required in navigation study to infer indirect meanings from stated facts. It is reasonable to suppose that an adequate measure of this ability would be obtained from a reading-comprehension test. The striking difference between pilot and navigator, in particular, as shown in table 5.1, promised one basis for discrimination of aptitudes and hence for classification.

Summary

To summarize, tests of verbal ability were incorporated in the Aviation Cadet Classification Program because of work done and results achieved with the verbal sections of the AAF Qualifying Examination, because of the hypothesis that individuals of high general and verbal intelligence are required to master the complexities of airplane operation and training, and because bombardier, navigator, and pilot differ in the requirements in this respect.

VOCABULARY TESTS

Cooperative Vocabulary Test, Form R, CI604A, CI605A

The Cooperative Vocabulary Test, Form R, was published by the Cooperative Test Service in 1941. It was included in the first classification battery during the winter and early spring of 1942.

Description.—The two code numbers, CI604A and CI605A, refer to the two different scores that were obtained from this test. CI604A refers to the *level*-of-comprehension score, whereas CI605A refers to the *speed*-of-comprehension score. The items are of a difficulty level appropriate for examinees with approximately 2 years of college training.

(1) *Internal characteristics.*—The test contains 210 items arranged in blocks of 30. The 30 items with the highest internal consistency are presented on the first page, the 30 with the next highest internal consistency are presented on the second page, and so on, in what is technically known as "cyclical construction." Items are not arranged according to difficulty.

(2) *Administration.*—The directions instruct the examinee to:

* * * Answer all the items you can on each page before going on to the next. Answer the items as they come: be careful not to skip pages. This is not a speed test, and your score does not depend as much on how many items you try to answer as it does on how many you get right on each page you attempt. On the other hand, the accuracy of your score will be decreased if you spend too much time on any one page. * * *

The time limit suggested by the publishers for many purposes is 30 minutes.

Following are two typical items. The examinee is asked to indicate which of the five alternative words is closest in meaning to the key word.

703320—47—5

Soothsayer:
1 Speaker.
2 Prophet.
3 Comforter.
4 Singer.
5 Peacemaker

Denote:
1 Regard.
2 Write.
3 Indicate.
4 Refuse.
5 Declare.

(3) *Scoring.*—As mentioned previously, two scores are obtained for each examinee. Both scores are obtained by application of the formula R—W/4 and the use of a conversion table yielding scaled scores.

Statistical results.—(1) *Distribution statistics.*—The distribution of level-of-comprehension scores in the test is indicated by a mean of 61.1 and a standard deviation of 9.8, based on a sample of 225 unclassified aviation students. For the speed-of-comprehension scores, a mean of 101.5 and a standard deviation of 41.9 was found for a sample of 243 unclassified aviation students. These two samples were tested on March 2 and March 4, 1942, at Psychological Research Unit No. 1.

(2) *Reliability.*—A reliability coefficient of 0.83 was obtained by the test-retest method for a sample of 438 pilot eliminees. This sample was tested in May and June, 1942, at Psychological Research Unit No. 3, and was retested 2 or 3 months later.

(3) *Test validity.*—Validation data are presented in tables 5.2 and 5.3.

TABLE 5.2.—*Validity Data for Cooperative Vocabulary Test, Form R, CI604A (Level)*

Group	Criterion	N_t	p_q	M_p	M_q	SD_t	r_{bis}
Pilots in primary training[1]	Graduation elimination	547	0.60	[2]6.14	[2]6.27	[2]1.74	−0.06
Bombardiers[3]	Record circular error	238	....	....	....	9.7	[4].02
Bombardiers[6]	Record circular error	194	....	....	....	7.9	[4]−.12
Navigators[3]	Graduation elimination	194	.88	64.5	64.5	9.2	.00
Navigators[7]	Graduation elimination	228	.79	61.5	64.2	9.0	.02
Navigators[8]	Graduation elimination	183	.84	65.5	60.3	9.0	.32

[1] Tested in April 1942 at Psychological Research Unit No. 3.
[2] In terms of scaled scores with a mean of 5.00 and a standard deviation of 2.00.
[3] Tested in the period April through August 1942 at Psychological Research Unit No. 3.
[4] Product-moment correlations.
[5] Record circular error is an unreliable criterion. Various estimates of its reliability vary between 0.00 to 0.40.
[6] Tested in the period February through April 1942 at Psychological Research Unit No. 1.
[7] In classes 42-10 to 42-17, Southeast Training Center. Tested in the period February through April 1942 at Psychological Research Unit No. 1.
[8] Reclassified pilots. For testing data and classes see footnote 7.

TABLE 5.3.—*Validity data for Cooperative Vocabulary Test, Form R, CI605A (Speed)*

Group	Criterion	N_t	p_q	M_p	M_q	SD_t	r_{bis}
Pilots in primary training[1]	Graduation elimination	547	0.60	93.45	97.50	26.85	−.09
Navigators[2]	Graduation elimination	194	.85	109.1	102.1	34.2	.11
Bombardiers[2]	Record circular error	237	....	.....	.....	35.5	[3].01

[1] Tested in April 1942 at Psychological Research Unit No. 3.
[2] Tested April through August 14, 1942 at Psychological Research Unit No. 3. Includes new aviation cadets and reclassified pilots.
[3] Pearson product moment correlation.

Evaluation.—From tables 5.2 and 5.3 it appears that vocabulary tests have slightly negative validity for pilots (for both level and speed scores).

Bombardier validity is approximately zero, though this conclusion must be taken with reservation, due to the unreliability of the criterion. The vocabulary test offered promise only as a navigator selection instrument, and even navigator validities were rather low and variable in these small samples. It may be concluded that within the range of ability of aviation students, verbal intelligence is not a factor for success in training except to a small extent for navigation. As an incidental comment, it may be said that the intercorrelation of speed and level scores is quite high ($r = 0.84$).

Vocabulary Test (AAF), CI604B

This test replaced Cooperative Vocabulary Test, Form R, CI604A, CI605A, in the classification battery during the late spring and summer of 1942. The Vocabulary Test (AAF) was a reprint of a commercial test prepared by the Cooperative Test Service.

Description.—The Vocabulary Test, AAF, contains words of appropriate difficulty for men with approximately 2 years of college training Thus, the difficulty is comparable to the preceding Cooperative Vocabulary Test, Form R.

(1) *Internal characteristics.*—The test contains 150 items, constructed in blocks of 30. The technique of cyclical construction is also employed here. The first 30 items are those with the highest internal consistency, the next 30 items are those with the next highest internal consistency, and so on. The words are not arranged in order of difficulty.

(2) *Administration.*—The time limit is 15 minutes. Approximately 3 minutes are required for the simple directions which specify that, "* * * this is not a speed test. Your score does not depend so much on how many items you try to answer as it does on how many you get right on each page you attempt." Answers are marked directly on a standard five-place IBM answer sheet.

(3) *Scoring.*—In spite of the admonitions in the directions to the contrary, the Vocabulary Test, AAF, is scored on a speed basis, in that only the speed-of-comprehension score is obtained. The scoring formula is R—W/4.

Statistical results.—(1) *Distribution statistics.*—For a sample of 1,000 unclassified aviation students (tested at Psychological Research Unit No. 3 in October and November 1943), the mean score was 48.1 and the standard deviation 19.9.

(2) *Internal consistency.*—The internal consistency of 81 of the items is indicated by a mean phi of 0.33, with a range from −0.12 to +0.88, and a standard deviation of 0.20, based on the highest and lowest 27% of 200 unclassified aviation students.

(3) *Difficulty.*—The difficulty level of items is indicated by a mean proportion of correct responses equal to 0.67, corrected for chance suc-

cess. The proportions ranged from 0.00 to 1.00 with a standard deviation of 0.27.

(4) *Factorial composition.*—In one analysis, the only significant loading appeared in the verbal factor, which this test helped to define. The loading was 0.71. It is practically a pure test for the verbal factor, with about 50% of its variance so allocated.

(5) *Test validity.*—Validity data are given in table 5.4.

TABLE 5.4.—*Validity data for Vocabulary Test, AAF, CI604B*

Group	Criterion	N_t	P_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$
Pilots in primary training[2]	Graduation-elimination	528	0.87	39.98	46.34	20.04	−0.17	−0.14
Pilots in primary training	Graduation-elimination	2,658	.73	.....	.....	.....	−.09	......
Pilots through basic training	Graduation elimination	1,942	.89	.....	.....	.....	.00	......
Bombardiers	Record circular error	320	...	.....	.....	22.9	[3] −.03	......
Navigators[4]	Graduation elimination	171	.94	69.7	62.3	23.9	.15	......

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In classes 44-H, 44-I, and 44-J. Tested at Psychological Research Unit No. 3.
[3] Product-moment r.
[4] Classified in the period Apr. 1 to Aug. 14, 1942 at Psychological Research Unit No. 3. Includes new aviation cadets and eliminated pilots.

Evaluation.—The Vocabulary Test, AAF (CI604B), is adequate, in terms of level of difficulty and reliability, for use in classification of air crew. It has a definite contribution to make to the selection of navigators, none for bombardiers, and for pilots it might well carry a small negative weight. One might question the wisdom of adverse selection of pilots on any trait as important as verbal intelligence, however. Even if such selection did improve graduation rate in training, it might work against selection of potentially superior plane commanders.

In a battery where measurement of the verbal factor is required, a vocabulary test is strongly to be recommended. Taken by itself, it is not as valid for selecting navigators as other, impure, verbal tests such as reading comprehension. Where uniqueness is a requirement, however, the vocabulary test has no rivals for the purpose of assessing verbal-comprehension ability.

READING COMPREHENSION TESTS

Reading Comprehension (Training and Duties), CI606A [2]

This is the first form of the Reading Comprehension Test in the classification battery. The paragraphs and questions concern the training and duties of the navigator, pilot, and bombardier for a special reason. In the early months of 1942, the roles of the navigator and bombardier in the air crew had not been extensively publicized. Consequently, most of the examinees were familiar only with the pilot's job. Few examinees were indicating first preference for navigator or bombardier training. It was felt that if information about all the air-crew positions were presented through the medium of this test, the number of stated preferences for

[2] Developed at Office of the Air Surgeon, Headquarters, AAF. Chief contributor: Lt. Col. Laurance F. Shaffer.

navigator and bombardier training would increase. Therefore, the training and duties test was administered before the preference blank (see chap. 26), on which examinees declared their first, second, and third choices for types of air-crew training. It was believed that if examinees were tested on the paragraphs, they would become more highly aware of their content. It was also believed that a verbal-comprehension score would be useful in selecting navigation students.

Description.—The training-and-duties paragraphs are simple descriptions of jobs, the training involved, and the individual characteristics required for success. The attempt to glamorize the roles of navigator and bombardier, while playing down the pilot, is obvious in the construction of the paragraphs. Following are some of the things said about the navigator and his job:

> The aviation cadet who is to become a navigator embarks upon a career involving the most modern application of one of the oldest of all the sciences * * * The extensive bombing experience of the present war has made clear the extremely difficult and important task of the navigator. It is his job, by day or by night, to chart the course that the bomber must fly from its base to the objective to be bombed and back to the home base. The navigator must be a person of superior intelligence, with a passion for accuracy and the power of logical reasoning under conditions requiring speed, coolness, and precision.

The bombardier, likewise, is played up as an extremely important member of the crew.

> The military value of a bombing plane is no greater than the ability of its bombardier to place his bombs on a military target. The bombardier should have as much intelligence as any member of the crew, and should possess unusual maturity of judgment and ability to accept responsibility; in addition, he must have the best of vision to pick out his target from great height; he must have superior muscle coordination to make delicate adjustments on the bomb sight, and he must remain calm and steady under combat conditions.

(1) *Internal characteristics.*—The test contains 30 scored items, 10 devoted to each job description. Most items are extremely easy and seem to serve the purpose of emphasizing the ideas presented in the paragraphs as well as that of testing.

(2) *Administration.*—The instructions for this test are simple and are all contained in the test booklet. Examinees are told to "Base your answers on the reading material or on inferences which can be drawn from it."

Thirty minutes are allowed for completion of the test.

Following are two typical items:

The member of the air crew who must be most apt in mathematics is:

A. The bombardier.
B. The navigator.
C. The pilot.
D. The radio operator.
E. The gunner.

The major offensive force of the air arm is:

A. The bomber plane.
B. The fighter plane.
C. The interceptor plane.
D. The observation plane.
E. The pursuit plane.

(3) *Scoring.*—The scoring formula is R—W/4.

Statistical results.—Fundamental data are quite complete on this test but with relatively small samples. All the data given below are for examinees tested in March 1942 at Psychological Research Unit No. 1.

(1) *Distribution statistics.*—The distribution of scores in this test is indicated by a mean score of 19.8 and a standard deviation of 5.1, based on a sample of 135 unclassified aviation students. The distribution was markedly negatively skewed.

(2) *Internal consistency.*—The internal consistency of items is indicated by a mean phi of 0.28 with a range from 0.00 to 0.60 and a standard deviation of 0.15, based on the highest 27% and the lowest 27% of 200 unclassified aviation students.

(3) *Reliability coefficient.*—A reliability coefficient of 0.86, corrected, was obtained by the odd-even method on a sample of 135 unclassified aviation students.

(4) *Difficulty.*—The difficulty level of items in the test is indicated by the mean proportion of correct responses equal to 0.80, corrected for chance success. The proportions ranged from 0.47 to 1.00 with a standard deviation of 0.18. These statistics are based on the data for 200 unclassified students.

(5) *Test validity.*—Validity data are shown in table 5.5.

TABLE 5.5.—*Validity Data for Reading Comprehension, C1606A*

Group	Criterion	N_t	P_p	M_p	M_e	SD_t	r_{bis}
Pilots in primary training	Graduation elimination	547	0.60	[1]6.00	[1]5.80	1.92	0.06
Bombardiers[2]	Record circular error	238	....	.....	.....	3.8	[3].00
Navigators[2]	Graduation elimination	193	.38	25.7	23.3	4.0	.31

[1] In terms of scaled scores with a mean of 5.00 and a standard deviation of 2.00.
[2] New aviation cadets and reclassified pilots, classified at Psychological Research Unit No. 3 from Apr. 1 to Aug. 14, 1942.
[3] Product-moment correlation.

Evaluation.—In view of the fact that the first form of Reading Comprehension, C1605A, was developed primarily to increase the number of preferences for navigator and bombardier training and was clearly deficient with regard to difficulty level, the statistical data are not very revealing. The easiness of the items is indicated by the unusually low difficulty level (0.80, corrected for chance). The test does not appear to be valid for the pilot training criterion or the bombardier circular error criterion. Its navigator validity on one small sample is fairly satisfactory.

Reading Comprehension Test, CI614C [3]

This is the seventh revision of a Reading Comprehension Test based on

[3] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Lloyd G. Humphreys, Maj. Merrill F. Roff, and Lt. Mahlon B. Smith.

material in two test booklets prepared in the Office of the Air Surgeon in June 1942. In December 1942, it was placed in the classification battery.

Description.—The paragraphs and the questions and answers used were carefully selected through item analyses of a large amount of trial material (the six previous forms). Two of the paragraphs were composed of material that might logically appeal to pilots, two others to navigators, and two others to bombardiers. The first paragraph deals with the subject of the rotating gun turret. The second paragraph describes the thrust and torque forces resulting from movement of the constant-speed propeller. Paragraph three discusses the north celestial pole and its relationship to various geographical positions on the earth's surface. Paragraph four describes the reasons for the drift of a projectile. The fifth paragraph involves a description and evaluation of the Mercator projection. The sixth and final paragraph tells of the formation and control of carburetor ice.

(1) *Internal characteristics.*—This test contains 30 scored items based on 6 paragraphs. Four to six items pertain to each paragraph.

(2) *Administration.*—The directions specify that 30 minutes are allowed for completion of the test. The administrator gives a time warning at the end of 10 minutes and again at 20 minutes. Answers are marked on the standard five-place IBM answer sheet. Following are two typical items, each pertaining to a different paragraph:

The turret always moves:

A. 360 degrees.
B. At an increasing speed.
C. In a circular path.
D. When the hand crank is turned.
E. When the clutch level is in the down position.

How does a Mercator projection, as compared to a globe, change the relative sizes of Norway and Spain?

A. There is no change in size since the Mercator projection is conformal.
B. Only the relative length of Norway is increased.
C. Only the relative width of Spain is increased.
D. Only the relative width of Spain is decreased.
E. The relative area of Norway is increased.

(3) *Scoring.*—The formula used in scoring Reading Comprehension CI614G is 2R—W/2, which is equivalent to R—W/4. Empirical studies of the optimal weight for W when the weight for R is unity, results in the conclusion that the formula R—W/7 is best for pilot selection. In samples of 1,096 and 1,226 pilots in primary training the empirical weights are −0.144 and −0.151, respectively. The validities for pilots to be expected from this formula yielded gains of only 0.001 in both instances, however, so no change in the traditional formula is called for. No corresponding data for navigators are available. Since the test is primarily a navigator-selection instrument, any modification of scoring

formula should be made in accordance with similar studies for that specialty.

Statistical results.—(1) *Distribution statistics.*—The distribution of scores in this test is indicated by a mean score of 19.3 and a standard deviation of 11.9, based on a sample of 1,095 unclassified aviation students, tested with the December 1942 Classification Battery, at Psychological Research Unit No. 1.

(2) *Internal consistency.*—The internal consistency of items is indicated by a mean phi of 0.43, with a range from 0.24 to 0.73, and a stand-

TABLE 5.6.—*Validity data for Reading Comprehension Test, CI1614G*

Group	Criterion	N_t	p_g	M_g	M_e	SD_t	r_{bis}	${}_c r_{bis}$[1]
Pilots in primary training[2]	Graduation-elimination	4,779	0.88	22.87	19.08	11.13	0.18	0.20
Pilots through basic training[3]	Graduation-elimination	3,046	.57	19.91	16.79	11.12	.18	.23
Pilots in advanced single engine training[4]	Average daily grades[5] .	277	...	.....	.....	.....	[6].03	.07
Pilots in advanced twin-engine training[7]	Average daily grades[5] .	360	...	.....	.....	.....	[6].00	.02
Pilots in B-17 transitional training[8]	Graduation-elimination	1,046	.98	21.20	15.88	11.87	.18	...
Pilots in B-24 transitional training[8]	Graduation-elimination	983	.92	20.80	17.51	11.62	.14	...
Pilots in B-25 transitional training[8]	Graduation-elimination	313	.98	21.39	23.71	11.66	−.08	...
Pilots in B-26 transitional training[8]	Graduation-elimination	380	.82	22.76	18.55	11.55	.18	...
Bombardiers[9]	Graduation-elimination	1,829	.79	16.51	14.79	10.26	.10	.13
Navigators[10]	Graduation-elimination	732	.87	24.10	17.12	11.76	.32	.47
Navigators[11]	Grades in dead reckoning (ground school)	463	...	.....	.....	.....	[12].28	.42
Navigators[11]	Grades in celestial navigation (ground school)	463	...	.....	.....	.....	[12].19	.31
Navigators[11]	Grades in dead reckoning (flight) .	463	...	.....	.....	.....	[12].09	.18
Navigators[11]	Grades in celestial navigation (flight)	463	...	.....	.....	.....	[12].18	.29
Navigators[11]	Grades in meteorology	463	...	.....	.....	.....	[12].30	.41
Navigators[11]	Military grades	463	...	.....	.....	.....	[12].10	.16
Navigators[11]	Final Composite Grades	463	...	.....	.....	.....	[12].26	.41
Radio-operator mechanics in training[13]	Graduation-elimination	235	.65	15.75	17.09	10.70	−.08	...
	Composite grades	153	...	.....	.....	.....	[12].39	...
Radio-operator mechanics in training[13]	Final academic grades	232	...	.....	.....	11.16	[12].41	...
Air mechanics in training								
Gunners in training[14] .	Air-to-air firing, percent hits[15]	194	...	.....	.....	11.6	[12].00	...

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44-F. Tested at Psychological Research Unit Nos. 1, 2, and 3.
[3] In class 43 J. Tested at Psychological Research Unit Nos. 1, 2, and 3.
[4] At Foster and Moore Fields. Tested between July 22, 1943 and Oct. 30, 1943 at Psychological Research Unit No. 2.
[5] Converted into normalized stanine scores.
[6] Z-averaged (average of coefficients for 2 schools) product-moment correlations.
[7] At Ellington and Frederick Fields. Tested between July 22, 1943 and Oct. 30, 1943 at Psychological Research Unit No. 2.
[8] In classes 43 J and 43 K. Tested at Psychological Research Units Nos. 1, 2, and 3.
[9] In classes 43-8, 43-9, 43-10, and 43-11. Tested at Psychological Research Units Nos. 1, 2, and 3.
[10] In classes 43-10 and 43-11. Tested at Psychological Research Units Nos. 1, and 2.
[11] In Hondo classes 43 10 through 43-15. 71 cases tested at Psychological Research Unit No. 1, 355 at Psychological Research Unit No. 2 and 37 at Psychological Research Unit No. 3.
[12] Product-moment correlations.
[13] Tested at Psychological Research Units Nos. 1, 2, and 3. Composite grades available only for graduates.
[14] In class 43-45. Tested at Psychological Research Unit Nos. 1, 2, and 3.
[15] This criterion was extremely unreliable.

ard deviation of 0.12, based on the highest 27% and the lowest 27% of 117 unclassified aviation students, tested in August 1942 at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—An odd-even reliability coefficient of 0.76, corrected, was obtained from a sample of 480 unclassified aviation students, tested at Psychological Research Unit No. 3.

(4) *Difficulty.*—The difficulty level of items in the test is indicated by the mean proportion of correct responses equal to 0.40, corrected for chance success. The proportions ranged from 0.00 to 0.84 with a standard deviation of 0.21. These data were based upon results from 117 unclassified aviation students, tested in August 1942 at Psychological Research Unit No. 3.

(5) *Factorial composition.*—Reading Comprehension, CI614G, was factor analyzed with four different batteries. It helped to define the verbal factor in each battery, with loadings of 0.54, 0.69, 0.65, and 0.65, and a weighted average of 0.60. There were significant loadings in the mechanical experience factor for the three analyses in which this factor appeared, the weighted-average loading being 0.37. The third highest loading for this test appeared in the general reasoning factor in all four analyses, the weighted average being 0.27. The loadings for the numerical factor in CI614G, as found in the four analyses, were 0.09, −0.02, 0.11, and 0.15 with a weighted average of 0.12, which contributed slightly to its validity for navigator selection, as did its reasoning variance. It will be seen in the discussion of the next form of this test (CI614H) how the numerical loading increased after an attempt was made to increase the navigator validity of the test.

(6) *Test validity.*—For validity data for various types of training, see table 5.6. For air-crew selection, this test has most validity for the navigator, next for the pilot at all levels of training, and lowest for bombardier. It has substantial promise for selection of radio operators and air mechanics.

(7) *Item validity.*—Items in this test were correlated with navigator training criteria (both preflight and advanced) and a bombardier-training criterion (preflight grades). The results are shown in table 5.7.

TABLE 5.7.—*Item Validity Data for Reading Comprehension, CI614G*

Group	Criterion	N_t	p_s	M phi	SD phi	Range of phi
Navigators[1]	Graduation-elimination	840	.72	0.04	0.06	−0.07 to 0.16
Bombardiers in preflight training[1]	Weighted average grade	190	...	.09	.10	− .21 to .28
Navigators in preflight training[1]	Weighted average grade	199	...	.28	.11	.03 to .47

[1] Tested in June 1943 at Psychological Research Unit No. 3.

Evaluation.—In CI614G, the seventh form of this test, a highly refined test of reading comprehension had been developed which could be considered an adequate measure of the ability to comprehend technical material. The hypothesis for the development of tests of verbal ability had held that such tests would be valid for all air-crew-officer positions. The validity statistics cited above, however, indicate wide difference in the amount of validity for pilot, bombardier, and navigator training success. It was evident that the greatest contribution that probably could be made by a reading-comprehension test was in the field of navigator selection. Further research was pointed towards an attempt to increase the navigator validity of the test.

Factorial results show that this is not by any means a pure verbal test. While 42% of its entire variance is verbal, about 14% must be allotted to mechanical experience, and about 6% to general reasoning. Its validity for the navigator may be due almost as much to its numerical component as to its verbal. The same may be said for its small bombardier validity. The reasoning component would contribute only to navigator validity.

Variations of the test.—Reading Comprehension, CI614G, was preceded by six preliminary forms. Since a test of this type is most satisfactory when there are several good questions for each paragraph of reading material, more time than usual was spent on trial runs and item analyses in order to maximize the number of differentiating items for the paragraphs selected. Difficulty arose not in finding good items, but in finding good complete sets of four to six items. The questions and alternate choices finally included were carefully selected on the basis of their effectiveness for the aviation student.

Editorial problems were ever present in constructing these preliminary forms. Paragraphs had to attain a certain difficulty, yet they had to possess a certain degree of clarity. Furthermore, it was desirable that they contain as much information as necessary for drawing direct or indirect inferences in answering items. In this respect, the revision was not quite successful, as the substantial mechanical variance shows. The variance indicates that even though it was believed that all necessary information had been provided, examinees still profit by previously gained mechanical experience in responding to items. The restrictions listed in this paragraph posed literary and semantic difficulties that were not easy to overcome.

Reading Comprehension, CI614H [4]

Form H of Reading Comprehension was specifically designed to discriminate between successful and unsuccessful navigators and is therefore more difficult than the previous forms designed to rank the entire aviation-student population.

Description.—(1) *Internal characteristics.*—This test is composed of 8 paragraphs concerning which 36 questions are asked. The items were designed to test ability to make valid inferences from the reading material as well as ability to answer more direct questions about the content. Short, succinct paragraphs were chosen, both because they lend themselves to

[4] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter, Jr., Capt. Frederick B. Davis, T/Sgt. Paul C. Davis, Lt. William M. Wheeler, and Lois G. Wright.

inferential items better than longer, more explicit passages, and because the number of items answered in the same length of time could be increased, thus increasing reliability. In order to prevent an increase in the correlations of the new form with other tests having heavy weights for pilot or navigator, items were selected on a basis of their lack of mechanical and numerical content as well as for their consistency with the total test. Thus, an item that had a high internal-consistency phi coefficient, and that was also correlated with the total score on one of the mechanical or numerical tests, was not so acceptable as an item with a similar phi that was not correlated with scores on the other tests. Paragraphs were taken from a wide range of technical material, including tests on navigation, physics, map reading, astronomy, and airplane instruments. This selection was made on the basis not only of greater pertinence to the type of reading the cadets would encounter in ground-school courses, but also of greater face validity. Material pertaining to all types of air-crew operation was included to make it appear that the test was one that pilots as well as others should take seriously.

A typical paragraph and the three questions asked about it are here reproduced to illustrate the inference-drawing technique.

Force and counterforce are equal and opposite. A force is always accompanied by a counterforce. The force on any one body is always exerted by some other body; this other body itself experiences an equal and opposite force. The two are parts, or different aspects, of one inseparable whole.

The general principle most justifiably derived from this paragraph is that:

A. The resultant of two forces cannot equal one opposing force.
B. All forces in the universe maintain equilibrium.
C. Force and counterforce differ in magnitude rather than direction and can thus be considered as parts of one inseparable whole.
D. The effect of a number of forces and counterforces acting on an object is always movement.
E. The work done by a body is not a function of the strength of the forces and counterforces.

In the case of pressing one's hand against a wall, the statements in the paragraph:

A. Would be true, depending upon the amount of force with which one pushed.
B. Would be true only if the wall were rigid.
C. Would be true only if the wall moved in the direction of the force.
D. Would be true only if something were pushing against the wall from the other side.
E. Apply without qualification.

The statements in the paragraph imply that a strong man striking a much weaker one would encounter a force:

A. Equal to his own.
B. Less than his own.
C. Greater than his own.
D. That would vary with the difference in strength between the strong and the weak man.
E. That would cause equal movement in the bodies of the weaker and the stronger man.

(2) *Administration.*—The directions for this sort of test were extremely simple, since the task was obvious. It was necessary only to caution the students to answer the questions on the basis of information contained in the paragraph. They were permitted to reread part or all of the paragraph while answering the questions. The time allowed to complete the test was limited to 30 minutes so that about one-third of the group would be able to finish.

(3) *Scoring.*—The standard scoring formula for five-choice items, multiplied by a factor of 2, was used; i. e., 2R—W/2.

Statistical results.—(1) *Distribution statistics.*—The data for several samples are given in table 5.8.

TABLE 5.8.—*Distribution Data for Reading Comprehension, CI614H*

Group	N	M	SD
Unclassified aviation students (post-college)[1]	1,500	20.8	12.4
Classified pilots[2]	1,676	19.3	11.3
West Point class 1946, classified pilots	888	33.3	14.5

[1] Tested November 1943 at Psychological Research Units Nos. 1, 2, and 3.
[2] In class 44-I. Tested at Psychological Research Units Nos. 1, 2, and 3.

TABLE 5.9.—*Reliability coefficients for reading Comprehension, CI614H, based upon groups of unclassified aviation students*

N	Type	r'	r_{tt}
1,000[1]	Odd-even	0.74	0.85
500[2]	Equated halves[3]	.52	.68

[1] Tested in April 1944 at Medical and Psychological Examining Unit No. 7.
[2] Tested at Medical and Psychological Examining Unit No. 10.
[3] Items assigned into two groups judged to be more or less equivalent.

(2) *Internal consistency.*—The internal consistency of items is indicated by a mean phi of 0.41, with a range from 0.13 to 0.63 and a standard deviation of 0.10, based on the highest 27% and the lowest 27% of 400 unclassified aviation students, tested in October 1943 at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—Reliability estimates are shown in table 5.9.

(4) *Difficulty.*—The difficulty level of items is indicated by the mean proportion of correct responses equal to 0.32, corrected for chance success. The proportions ranged from 0.01 to 0.66 with a standard deviation of 0.16. The sample consisted of the 400 unclassified aviation students tested in October 1943 at Psychological Research Unit No. 3.

(5) *Factorial composition.*—Reading Comprehension, CI614H, was factor analyzed with two different batteries. Contrary to expectations, the loading in the verbal factor did not increase over that in the G form of the test. The verbal loadings for the H form were 0.58 and 0.59 as compared to a weighted average loading of 0.60 for the G form. The loading in the [illegible] experience factor did decrease as had been desired, although the amount is uncertain. The loadings were 0.33 and 0.04 in two analyses as compared to a weighted average of 0.37 for the three G-form

analyses. It had been desired that the numerical content of form H would not be greater than that of form G, to prevent an increase in the correlations of the new form with other tests heavily weighted for navigator. An attempt had been made, as mentioned previously, to select items partially on the basis of their lack of mechanical and numerical content as well as their consistency with the total test. It was not possible, however, in seeking complex reading material, to find paragraphs entirely free of mechanical or numerical content. This fact was emphasized by the characteristic loading of 0.14 in the numerical factor.

(6) *Test validity.*—There are validity data for air mechanics as well as for pilot training, but unfortunately none for the navigator (see table 5.10).

TABLE 5.10.—*Validity Data for Reading Comprehension, CI614H*

Group	Criterion	N_t	p_s	M_s	M_e	SD_t	r_{bis}	ρr_{bis}
Pilots in primary training[1]	Graduation-elimination ..	1,676	0.89	19.46	18.24	11.29	0.06	[2]0.14
Pilots in primary training[3]	Graduation-elimination ..	3,145	.84	21.03	18.60	12.47	.11	[4].18
Air mechanics in training[5]	Final average grade	254	....	.....	.....	.....	[6].28	.....
Air mechanics in training[5]	Final average grade	428	....	.....	.....	8.22	[6].20	.....
Armament trainees[7]	Average armament grades	269	....	.....	.....	10.65	[6].22	.....

[1] In class 44J, tested at Psychological Research Units Nos. 1, 2, and 3.
[2] Assuming an unrestricted augmented stanine standard deviation of 1.91.
[3] In class 44I, tested at Psychological Research Units Nos. 1, 2, and 3.
[4] Assuming an unrestricted stanine standard deviation of 2.00.
[5] Tested with the November 1943 Classification Battery at Medical and Psychological Examining Unit No. 6.
[6] Product-moment correlation.
[7] In Lowry Field armament classes 34–44A and 35–44A.

Evaluation.—Examination of validity data obtained on Reading Comprehension, CI614G, indicated that an attempt should be made in the next form (CI614H) to increase the validity of the test for navigator success. Form H, then, was designed specifically to separate good from poor and mediocre navigator material without regard for discrimination at the lower levels, but with regard for ranking the more apt candidates. The test, therefore, is more difficult than the previous forms designed to rank the entire aviation-student population. It was planned to revise Form G, increasing the verbal factor content, decreasing the mechanical experience content, and at least holding constant the correlation of reading comprehension with numerical tests. As pointed out in the discussion of factor content, the loading in the verbal factor did not increase in form H. The loading in the mechanical experience factor did decrease, although certainly not as much as had been hoped, and the usual loading appeared in the numerical factor. No validity statistics have been computed for navigators on Form H, but the factor results indicate that the goals anticipated in the CI614H revision were not realized. A further revision is therefore in order.

Variations of the test.—This form was preceded by five preliminary forms, CI614HX1, HX2, HX3, HX4, and HX5. Form HX1 contained

38 items based on 4 paragraphs, on the subjects of radio beams, the function of an air-speed meter, parallax, and the magnetic pole in the Northern Hemisphere and its properties. Form HX2 contained 38 items based on 6 paragraphs, all different from HX1. These paragraphs discussed (1) Bridgman's work in modern physics, (2) interaction of waves and mass, (3) force and counterforce, (4) navigational bearing, (5) radio bearing, and (6) compensation of a compass. Except for the last, these paragraphs were considerably shorter than those in the first preliminary form. The next revision (HX3) incorporated three paragraphs from HX1 and one form HX2, with 28 items. Form HX4 clarified and shortened the paragraphs on Bridgman's contribution to modern physics, force and counterforce, and compass compensation. To these were added paragraphs on sound waves and spectrum colors. Twenty-eight items were based on these five paragraphs.

The final experimental form, HX5, was lengthened to include 9 paragraphs and 50 items. The testing was increased to 50 minutes, with a provision for extra time, if necessary, to allow 75 to 80% of the students to finish. The paragraphs as they appeared in this form were rewritten, where this was deemed desirable, on the basis of internal-consistency-item analysis of the previous forms. In the classification battery form, CI604H, the test was cut to 36 items and 8 paragraphs. Those included were on (1) the magnetic North Pole, (2) force and counterforce, (3) spectrum colors, (4) compass compensation, (5) Mercator projection, (6) air-speed meter, (7) bearing, and (8) Bridgman's work in physics.

After Reading Comprehension, CI614H, was placed in the classification battery, an alternate form, CI614JX1, was constructed. The type of reading material included in the paragraphs is similar to Form H, although the actual subject matter of all the paragraphs in JX1 is different. JX1 contains 42 items based on 7 paragraphs. Forty minutes are allowed for completion of the test. The paragraphs include discussions of (1) radio television, (2) altitude tolerance, (3) supersonics, (4) hydraulic systems, (5) atmospheric pressure, (6) refraction of light rays, and (7) steam turbines.

Psychomotor Instruction Comprehension Test, CI626B [6]

This test was designed to measure the comprehension of instructions given for psychomotor classification-battery tests.

Description.—The test is administered after students have completed the six psychomotor tests. If the administration takes place immediately following the last psychomotor test, memory factors are minimized. This is desirable, since the test is designed primarily as a measure of comprehension. Diagrams of each psychomotor test are presented, and the examinee is asked questions about each task.

[6] Developed at Medical and Psychological Examining Unit No. 10. Chief contributor: Capt. Joseph E. King. An earlier test of the same type was developed at MPEU No. 9. Chief contributor: S/Sgt. Arthur Z. Cerf.

(1) *Internal characteristics.*—The test contains 83 items, 6 of which are on the instructions given preliminary to the psychomotor tests, 61 are on specific tests, 5 are on conditions for making good scores, and the remaining 11 on distinguishing features of the tests.

(2) *Administration.*—Answers are marked directly on the standard five-place IBM answer sheet. As mentioned previously, it is imperative that this test be administered immediately after completion of the psychomotor battery. Two sample items are duplicated below. The first refers to figure 5.1, the second to figure 5.2.

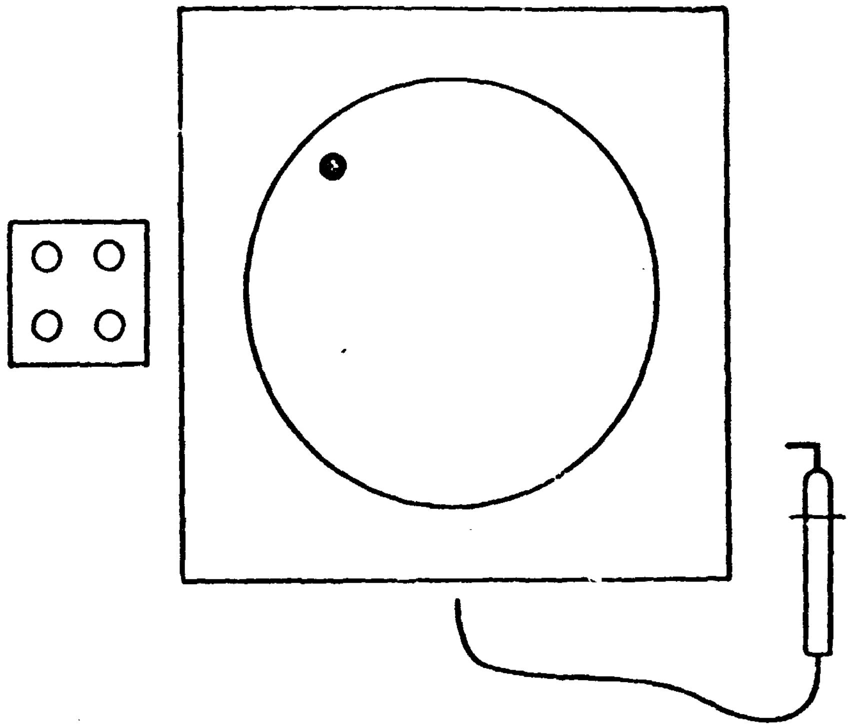


FIGURE 5.1
SAMPLE ITEM OF PSYCHOMOTOR INSTRUCTION COMPREHENSION, CI626B

Your task in this test was to:

A. Follow a moving target.
B. Keep a stylus level during each trial.
C. Use a smooth, free-swinging motion of the arm and shoulder while following the moving target.
D. Keep the end of the stylus on the brass target as it moved.
E. Keep the stylus one inch off the target as it moved.

(3) *Scoring.*—The score on this test is simply the number of correct responses.

Statistical results. (1) *Distribution statistics.*—The distribution of scores is indicated by a mean score of 58.90 and a standard deviation of 7.4, based upon a sample of 400 unclassified aviation students.

(2) *Internal consistency.*—The internal consistency of items is indicated by a mean phi of 0.26 with a range from 0.00 to 0.59 and a stand-

ard deviation of 0.11, based on the highest 27% and the lowest 27% of 400 unclassified aviation students.

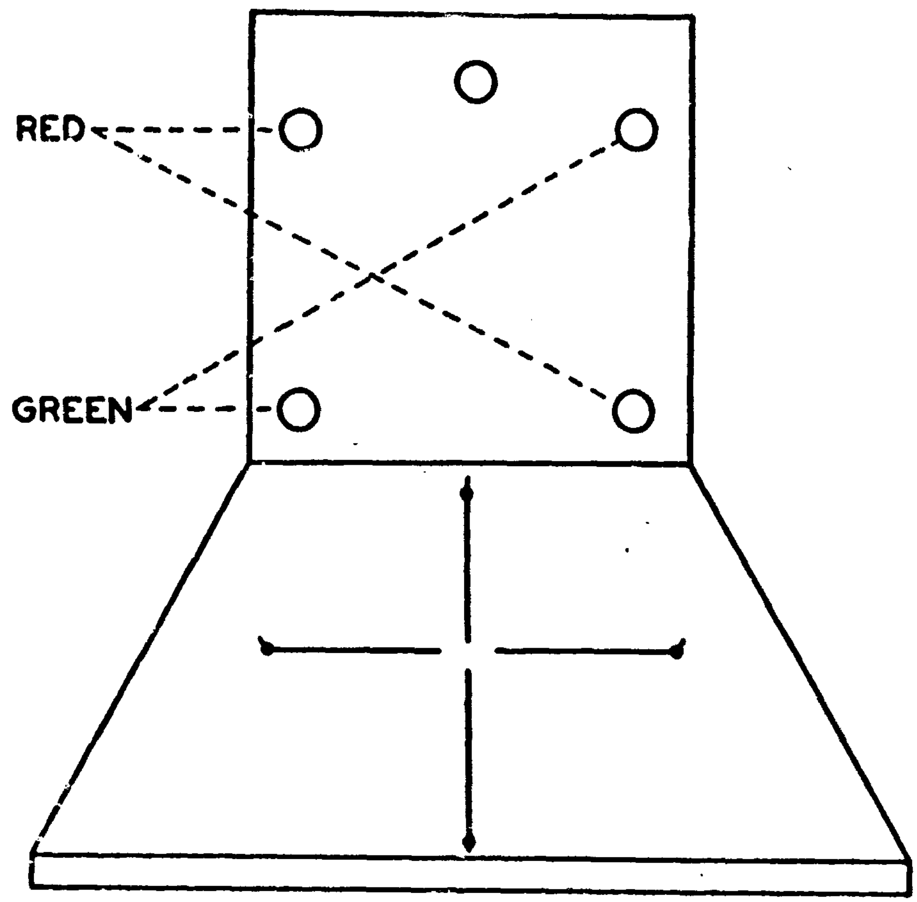


FIGURE 5.2
SAMPLE ITEM OF PSYCHOMOTOR INSTRUCTION COMPREHENSION, C1626B

If the apparatus showed this pattern, you were to snap the:

A. Lower switch.
B. Upper and lower switches.
C. Right switch.
D. Upper switch.
E. Left and right switches.

(3) *Reliability coefficient.*—A reliability coefficient of 0.75, corrected, was obtained by the odd-even method on a sample of 400 unclassified aviation students.

(4) *Difficulty.*—The difficulty level of items in this test is indicated by a mean proportion of correct responses equal to 0.71, corrected for chance success. The proportions ranged from 0.00 to 0.99 with a standard deviation of 0.22. These data are based upon results from 400 unclassified aviation students.

Variations of the test.—Psychomotor Instruction Comprehension Test, C1626B, was preceded by a preliminary form, C1626A. This contained

30 items based on the 6 classification-battery psychomotor tests. Five items were based on the orientation talk, an introductory speech given the examinees before testing began. The chief difference between the preliminary and final forms of this test is that in the earlier variation, CI626A, all the items were completely verbal.

Evaluation.—This type of test has possibilities as a measure of ability to remember instructions and should be studied in connection with memory tests, particularly the test, Memory for Tactical Plans (see ch. 11). It probably has nothing to offer as a test of verbal comprehension. If a memory test is desired, one of purer composition and one with more satisfactorily controlled administration could probably be designed.

EVALUATIONS AND CONCLUSIONS

Vocabulary Tests

Statistical results have shown considerable differentiation between the two types of tests used to measure verbal intelligence, as far as their factorial composition and their ability to predict air-crew success is concerned. The failure of part of the hypothesis that verbal ability is valid for air-crew training success was indicated in the case of pilots and bombardiers by results from the vocabulary tests. These results revealed a slightly negative validity for pilots, a zero validity for bombardiers, and a small validity for navigators (approximately 0.20). Vocabulary is the best measure of the verbal factor, having a loading of 0.71. Although this factor has some validity for navigators, reading comprehension tests have a greater navigator validity than the vocabulary tests because of their reasoning and numerical components.

Owing to its limited predictive value, the vocabulary test was dropped from the classification battery in the summer of 1942. It was replaced by the Technical Vocabulary Test, CE505C, which is a test of specific technical information pertaining to piloting, navigation, and bombardiering. This test, although related to vocabulary, and loaded with the verbal factor (0.41 for the pilot score), possessed a validity of 0.21 for pilots. Part of this validity is derived from the test's loading with the mechanical experience factor (0.39). The remaining part is accounted for by its loading with pilot interest (0.34).

Reading Comprehension Tests

The Reading Comprehension test has proved to be a useful classification instrument, particularly for predicting navigator success. Mean validities of 0.38 for navigators, 0.13 for bombardiers, and 0.20 for pilots have been obtained. The pilot validity comes largely from a loading with the mechanical experience factor (0.37 for form G). Although the verbal factor is valid for the navigator, it is not the most important ability involved in the navigator criterion. Furthermore, other tests in the classification battery have covered this factor fairly well (General Informa-

703120—47—6

tion, navigator score, with a verbal loading of approximately 0.60, and Technical Vocabulary, navigator score, with a verbal loading of approximately 0.75). A test to be highly valid for navigators should contain a significant loading in the numerical factor. In spite of the navigator validity of Reading Comprehension, CI614G, the numerical factor loading of the test is low (0.14).

As far as future policy is concerned, it might be advisable to work toward removal of reading comprehension from the classification battery, in spite of its present contribution to the stanines. The reason is that reading comprehension is factorially complex. It has effective loadings in four factors—verbal, mechanical experience, numerical, and general reasoning. Reading Comprehension's communality is high (0.87), which indicates that almost all of the validities of the test a-- derived from the four identified valid factors. The four factors are adequately covered by other tests in the classification battery, so that Reading Comprehension is merely a duplicate measure of them. It might be profitable, therefore, to attempt to increase the loading in each of these factors for the particular test that is the best known measure of the factor. If this were done, the classification battery would contain purer measures of these factors. The function of Reading Comprehension would then be more than adequately supplanted.

BIBLIOGRAPHY

(1) Bingham, Walter V. *Aptitudes and Aptitude Testing,* New York: Harper and Brothers, 1937.

(2) *A Report on the Purpose, Development, and Validation of Test AC10A.* Washington, D. C.: Psychological Division, Office of the Air Surgeon, Headquarters of the Army Air Forces, 1942.

CHAPTER SIX

Mathematical Tests[1]

INTRODUCTION

Prediction of Academic Achievement

Measures of mathematical aptitude or achievement have long been used to assist in the prediction of success not only in more advanced mathematics but in other academic pursuits as well. The use of these measures to predict success had three principal bases. First, it was logical to suppose that performance in mathematics would be positively correlated with performance in pursuits having mathematical or numerical content. Second, mathematical ability was generally considered to be one of the best indices of abstract intelligence. Third, it was widely assumed that the exact methods and critical attitudes demanded in mathematics would be carried over into the performance of other tasks. With respect to the third point, it is generally agreed that much less transfer of training takes place than was once supposed. Even though not accepted unreservedly, these hypotheses offered a basis for believing that mathematics tests would prove useful in predicting success in air-crew training.

History of Mathematics in Air-Crew Selection

The early recognition of the importance of mathematics is attested by the place assigned to it in examinations administered to determine qualifications for air-crew training. As noted in chapter 4, examinations in lieu of high-school graduation and in lieu of completion of 2 years of college were initiated in 1920 and 1927 respectively. Mathematics was apparently prominent in both levels of examination. The standardized objective examination adopted in 1941 gave great emphasis to mathematics, four of the five required sections being mathematics (arithmetic, algebra, geometry, and trigonometry). In like manner, the first form of the AAF Qualifying Examination which was adopted in January 1942 had mathematics as an important constituent. One of the six sections was mathematical and contained three types of problems: arithmetic reasoning, numerical operations, and mathematics achievement. Significantly, performance on this section proved to have a biserial correlation of 0.64 with graduation-elimination from navigation school. This figure was based on a group of 174 graduates and 47 eliminees.

With the establishment of bombardier and navigator training schools in 1940, the question of classification for air-crew training arose. This problem was at first solved by selecting bombardiers and navigators from

[1] Written by T/Sgt. Paul C. Davis.

those eliminated from pilot training. Faculty boards that eliminated pilots decided the type of training to which such eliminees should be sent. Appreciation of the importance of mathematics in predicting success in navigation is demonstrated by the fact that the faculty boards attempted to send to navigator training men who were trained in engineering and mathematics. Recognition of the more intellectual or academic nature of navigation is also indicated by the fact that eliminees for failure in ground school were not sent to navigation training.

Job Analysis Findings

The early job investigations described in chapter 1 revealed important facts concerning job requirements and yielded leads for construction of predictive instruments. According to these analyses, mathematics had an important place in air-crew tasks.

The Navigator.—Even superficial examination of the duties involved in navigation reveals that relatively high degrees of skill in making computations and in interpreting data are required. In fact, early job analyses of the duties of navigators disclosed that numerical and mathematical abilities were probably the most important factors influencing success in navigation.

Among the many duties of the navigator that demand mathematical knowledge and skill, typical examples may be cited. The navigator must calculate drift, distance, and direction from data gathered from various instruments. He also uses computers, such as the E-6B,[3] accurate use of which requires considerable skill and mathematical knowledge. The navigator also uses many mathematical tables, both in making calculations (such as tables of squares and roots) and in looking up pertinent data. In addition to these specific abilities, the navigator must have a keen sense of the interrelationships of the facts which he has gathered and must be able to integrate knowledge of these facts into a clear and accurate picture which will enable him to make valid navigational decisions.

In one early survey, an analysis was made of reasons for failure of navigator trainees, as reported by members of a navigation school staff. A total of 56 responses was categorized according to cause of failure (related to intellectual as distinguished from physical and emotional causes). Of these, 37 indicated, either directly or indirectly, a lack of speed or ability in numerical or mathematical tasks.

Analyses also indicated that, apart from the distinctly mathematical phases of the navigation job, the general content of the task is much more academic than the jobs of either pilot or bombardier. The high correlation of mathematical ability with general academic achievement lent support

[3] The E-6B computer is a circular slide-rule and a device for solving vector problems encountered in dead reckoning. The slide-rule face may be used for solving problems involving time, speed, distance, multiplication, division, proportions, true air speed (from calibrated air speed), and true altitude (from calibrated altitude). The face used in solving vector problems has a transparent plotting disk with a graduated compass rose which can be rotated with the fingers. A slide under the disk is marked with concentric speed circles, radiating drift lines, and a rectangular grid. The slide is seen through the plotting disk and may be moved back and forth under the disk as desired. The plotting is done in pencil on the transparent disk.

to the hypothesis that mathematical ability would be a good predictor of success in navigation.

The pilot.—In general, there was less reason to expect high correlation between mathematical ability and success in pilot training than in navigation training, although early concepts of the requirements of the pilot's job pictured it as a fairly intellectual task. The qualifying requirement of 2 years of college training or its equivalent during the years 1927 to 1942 indicates the emphasis placed upon academic aptitude and achievement. If such standards be valid, mathematical ability could reasonably be expected to be positively correlated with success in pilot training.

Early pilot-job descriptions mention table reading, use of computers, and simple mathematical calculations. These functions are apparently incidental, however, and constitute a relatively small proportion of the pilot's job. It might be assumed justifiably, therefore, that any candidates who passed the preliminary hurdles to pilot training would be able to perform these mathematical tasks satisfactorily.

The bombardier.—The crucial test of the bombardier's proficiency comes during the "bomb run." The entire success of the bombing mission depends upon the speed and accuracy of his performance during the few seconds preceding the bomb release. In order to set proper data into the bomb sight, the bombardier must read several tables correctly, use a computer accurately, and make relatively simple calculations, e. g., interpolation. One of the most important and exacting of these specific duties is determining true altitude on the basis of temperature readings and other pertinent data. This computation is of special importance, because error will inevitably result in a short or over bomb drop unless there happen to be compensating errors.

Compared with the navigator, the bombardier has fewer mathematical data to integrate. The most important mathematical requirement of the bombardier is that he perform the necessary calculations speedily and accurately, since the time element in the bombing run makes decisions based on these calculations practically irrevocable.

Summary.—Job-analysis findings indicate that mathematics is extremely important to navigation and that measures of proficiency or achievement in mathematics should be good predictors of navigational success. To a much less degree, it appeared that mathematical ability affects success in bombardiering. For the pilot it appeared that little relationship exists between success and mathematical ability. In this chapter two types of tests will be discussed—general mathematics and numerical computations. Arithmetical-reasoning tests, although mathematical, are primarily reasoning tests and so will be discussed in chapter 7.

GENERAL MATHEMATICS TESTS

Mathematics A, CI702E [3]

This form is typical of those tests devised to measure ability and

[3] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Lloyd G. Humphreys, Maj. Merrill F. Roff.

achievement in advanced arithmetic, algebra, and trigonometry.

Description.—This test was designed to measure competence in mathematics. In general, a student who has completed high-school algebra should be equipped to solve most of the problems.

(1) *Internal characteristics.*—The test consists of 30 five-alternative, multiple-choice items. The last three items require knowledge of trigonometry. The following problems are typical of the first 27 items:

$(3x-1)\ (2x+2) =$

A. $6x^2-x+2$
B. $6x^2+4x-2$
C. $6x^2-4x-\frac{1}{2}$
D. $3x^2-x+3$
E. x^2+2x-6

$R = c^2d^3$. If $c = 2$ and $d = -3$, then $R =$

A. -108
B. -72
C. 36
D. 72
E. 108

If $S = 3MR^2$, then $M =$

A. $3R^2S$
B. $\frac{SR^2}{3}$
C. $\frac{3S}{R^2}$
D. $\frac{S}{3R^2}$
E. $\frac{3R^2}{S}$

(2) *Administration.*—The examinees are urged not to spend an undue amount of time on problems they find difficult. Scratch paper is furnished for any necessary written computations.

(3) *Scoring.*—The test was scored first with the formula $R—W/4$ and later with the formula $2R—W/2$. The change was made in order to obtain scores of a magnitude which better fitted the system of weighting in computing the composite classification score.

Statistical Results.—Quite complete statistical data were obtained, since this test was included in the classification battery for some time.

TABLE 6.1.—*Validity Data for Mathematics, CI702E, for pilots in various stages of training*

Group	Class	Psychological Research Unit No.	Criterion	N	p_e	M_g	M_e	SD_t	r_{bis}	r_{bis}[1]
In primary training	43J	1, 2 and 3	Graduation-elimination	3,151	0.66	18.12	16.48	13.55	0.07	0.12
In primary training	43F	3	Graduation-elimination	2,376	.63	18.76	17.08	13.37	.08	....
In primary training	43E	2	Graduation-elimination	1,148	.76	19.8	20.4	13.6	−.02	....
In primary training	43D	2	Graduation-elimination	1,520	.75	21.1	16.4	14.1	.20	....
In primary training	43F	3	Graduation-elimination	1,429	.84	19.31	17.18	13.27	.08	....
In basic training	43J	1, 2 and 3	Graduation-elimination	2,978	.54	18.4	16.7	13.6	.07	.12
Through advanced training	43J & 43K	1, 2 and 3	Graduation-elimination	1,046	.98	21.2	20.2	14.8	.03	....
In B-17 transitional training	43J & 43K	1, 2 and 3	Graduation-elimination	983	.92	20.4	18.1	14.6	.08	....
In B-24 transitional training	43J & 43K	1 and 2	Graduation-elimination	313	.98	20.1	21.4	14.3	−.04	....
In B-25 transitional training	43J & 43K	1, 2 and 3	Graduation-elimination	380	.82	21.4	15.9	15.1	.21	....
In B-26 transitional training	43J & 43K	1, 2 and 3	Graduation-elimination	2,415	.64	20.6	21.7	15.1	−.04	....
In P-40 assignment	43J	1	Total percent hits	459	....	.....	.....	.....	[2].11	....
In fixed gunnery training	43J	1	Qualifying percent hits	459	....	.....	.....	.....	[2].07	....
In fixed gunnery training	43K	1	Total percent hits	403	....	.....	.....	.....	[2].00	....
In fixed gunnery training	43K	1	Qualifying percent hits	403	....	.....	.....	.....	[2].00	....
In advanced single-engine training	43J-44A	3	Proficiency rating	562	....	.....	.....	.....	[2].07	....
In advanced twin-engine training	43J-44A	3	Proficiency rating	685	....	.....	.....	.....	[2].03	....

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Product-moment correlations.

(1) *Distribution statistics.*—Based on a sample of 9,622 unclassified aviation students (tested at Psychological Research Units Nos. 1, 2, and 3 with the June 1942 Classification Battery), the test yielded a mean score ($2R—W/2$) of 19.4 and a standard deviation of 14.1. The distribution curve for new aviation students, such as the sample cited above, was positively skewed and markedly flatter than normal.

(2) *Internal consistency.*—Internal-consistency item analysis of this test revealed a marked degree of homogeneity. Internal-consistency phi values based on administration to 400 unclassified aviation students ranged from 0.10 to 0.84, with a mean of 0.55 and a standard deviation of 0.19.

(3) *Reliability coefficients.*—Reliability of the test, estimated by the odd-even method, was 0.92 (corrected for length), based on 200 cases tested at Psychological Research Unit No. 2 in April 1942.

(4) *Difficulty.*—Based upon the item analysis previously referred to, the test yielded a mean difficulty index, corrected for chance, of 0.48 with a standard deviation of 0.18. Testing of a group of students who had been selected for superior performance in examinations taken prior to entering college training detachments (upper 20 percent based upon composite of achievement-test scores in geography, history, mathematics, physics, and reading comprehension) yielded a mean difficulty index, corrected for chance, of 0.68 and a standard deviation of 0.15.

(5) *Factorial composition.*—This form proved to have significant loadings with three factors only. The verbal factor had a loading of 0.53, the numerical factor 0.42, and the visualization factor 0.33, in an analysis in which a general-reasoning factor also appeared, but with a loading of only 0.12.

(6) *Test validity.*—Validation data were secured against all air-crew and some technical-specialty criteria. The data are shown in tables 6.1 to 6.4 inclusive.

TABLE 6.2.—*Validity data for Mathematics, CI702E for bombardiers in training*

Class	Research unit	Criterion	N_t	p_s	M_p	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
43-5 to 43-7[2]	1	Graduation-elimination	552	0.84	13.3	11.7	11.2	0.08	....
43-5 to 43-7[2]	2	Graduation-elimination	329	.86	22.0	20.9	14.4	.04	....
43-5 to 43-7[2]	3	Graduation-elimination	469	.82	20.3	18.9	12.8	.06	....
43-8 to 43-11[3]	1, 2, 3	Graduation-elimination	1,829	.79	18.3	15.8	12.4	.12	0.14
43-14 to 43-18[3]	1, 2, 3	Graduation-elimination	456	.84	21.3	15.5	16.2	.20	.23
43-14 to 43-18[4]	1, 2, 3	Graduation-elimination	524	.86	22.3	18.0	13.0	.18	.23
43-1 to 43-4	3	Average grades	195	....	....	....	....	[5].22	....
43-1 to 43-4	3	Record circular error[6]	195	....	....	....	....	[5]−.09	....
43-1 to 43-4	3	Combat circular error[6]	195	....	....	....	....	[5].08	....

[1] Assumed unrestricted stanine standard deviation not reported.
[2] New aviation cadets, taking 12-week course (no navigation training).
[3] New aviation cadets, taking 18-week course (with navigation training).
[4] Reclassified pilots, taking 18-week course.
[5] Product-moment correlations.
[6] A highly unreliable criterion.

TABLE 6.3.—*Validity data for Mathematics, CI702E, for navigators in training*

Class	Research unit	Criterion	N_t	p_p	M_p	M_e	SD_t	r_{bis}	cr_{bis}
43–10 and 43–11	1, 2	Graduation-elimination	728	0.87	35.07	29.36	13.87	0.22	[a]0.43
43–12 to 43–15	1, 2, 3	Graduation-elimination	1,969	.79	34.51	29.14	13.78	.22	[a].41
43–10 to 43–15[b]	1, 2, 3	Grades in dead reckoning (ground school) .	463	....	.....	.....	.....	.34	[c].49
43–10 to 43–15[b]	1, 2, 3	Grades in celestial navigation (ground)	463	....	.....	.....	.....	.24	[c].38
43–10 to 43–15[b]	1, 2, 3	Grades in dead reckoning (flight)	463	....	.....	.....	.....	.17	[c].26
43–10 to 43–15[b]	1, 2, 3	Grades in celestial navigation (flight)	463	....	.....	.....	.....	.18	[c].31
43–10 to 43–15[b]	1, 2, 3	Grades in meteorology	463	....	.....	.....	.....	.25	[c].38
43–10 to 43–15[b]	1, 2, 3	Military grades	463	....	.....	.....	.....	.11	[c].17
43–10 to 43–15[b]	1, 2, 3	Final composite grades	463	....	.....	.....	.....	.32	[c].48

[a] Assuming an unrestricted stanine standard deviation of 2.00.
[b] At Hondo Army Air Field.
[c] Product-moment r's. Assumed unrestricted stanine standard deviation not reported.

TABLE 6.4.—*Validity data for Mathematics, CI702E, for miscellaneous specialties*

Group	Criterion	N_t	p_s	M_s	M_f	SD_t	r_{bis}
Radio operator mechanics	Graduation-elimination	235	0.65	20.16	15.38	19.23	0.19
Air mechanic armorers[1]	Average grades	232	....	.....	.....	14.74	[2].21
Air mechanic armorers	Average grades	376	....	.....	.....	13.18	[2].03
Flexible gunners[3]	Air-to-air firing	194	....	.....	.....	13.3	[2].11
Flexible gunners[3]	Final examination	194	....	.....	.....	13.3	[2].06

[1] Tested at Psychological Research Unit No. 2. Entered training between March and July 1943 at AAF Technical School, Sheppard Field.
[2] Product-moment correlations.
[3] In classes 43-45 and 43-48, at Buckingham Army Air Field.

(7) *Item validity.*—Items of this test were validated against the pass-fail navigator criterion. The mean validity thus obtained was 0.09, the standard deviation was 0.09, and the range was from −0.04 to +0.33. Ten of the 30 items yielded validities of less than 0.05, which indicates that careful revision and selection of items might increase the over-all test validity.

Evaluation.—This test proved to be a relatively good predictor of success in navigation training. For other air-crew tasks and technical specialties which are not highly loaded intellectually, the test proved a much less satisfactory predictor.

Variations.—In developing and refining a general mathematics test, several successive forms were prepared and administered.

(1) General Mathematics—Form A[4].—This is the first form of general mathematics developed for use in classification. It contains 75 items, which are arranged in three parts of 25 items each. The parts are timed separately, 15 minutes for each part. The test contains arithmetic-reasoning items as well as mathematics items similar to those described under form CI702E. This form was used in the classification of air-crew candidates for a short time and was validated against the navigator pass-fail criterion in training. For a sample of 478 cases (tested at Psychological Research Unit No. 2), the test yielded a biserial correlation of 0.51 with navigator success. Although the test is too long and difficult, the results provided considerable impetus for further exploration of the usefulness of mathematics tests.

(2) General Mathematics Test, Form II (CI702B)[5].—This form is a revision of Form A made easier and shortened to 60 items. The same amount of time is allowed as for Form A. The same categories as those in Form A are retained, but some slight changes in the numbers of items in the various categories were made. The categories and their contents are: Algebraic-equations and formulas, 21 items; arithmetic, 16 items; plane geometry, 8 items; trigonometry, 5 items; algebra, 4 items; analytic geometry, 4 items; spherical geometry, 1 item; and solid geometry, 1 item.

The difficulty level of this form of general mathematics is much more appropriate than that of the previous form. Analysis revealed that some

[4] Developed at Psychological Research Unit No. 1. Chief contributor: Lt. Col. Laurance F. Shaffer.
[5] Developed at Psychological Research Unit No. 3 by the test-construction staff.

items were still extremely easy and some extremely difficult, while some had low internal consistency. This form was used for a short time only in classification.

(3) General Mathematics Test, Form III (CI702C)[e].—This form also contains three parts of 20 items each and is administered with a total time-limit of 45 minutes. It is a revision of form II in which the extremely hard items, the extremely easy items, and those with low internal consistency were either revised or replaced. Analysis was also made of the items in terms of the place where such material was covered in the navigation ground school. This form is less homogeneous (mean internal-consistency phi = 0.39) and somewhat more difficult (mean difficulty index, corrected for chance = 0.40) than form CI702E, previously described.

(4) Mathematics A (CI702F)[f].—This form is a revision of the CI702E form and contains 35 items, which is 5 more than in form E. The time allowed for the test is 25 minutes. This form proved, as expected, to be more difficult than the E form. The mean difficulty index for students who had just taken mathematics in the college training detachment is 0.46, as compared with a mean of 0.68 in Form E for a similar group. This form was in the classification battery for more than a year, as a navigator-selection instrument.

Factor analysis of this form of the test revealed considerable difference from form E, although the contents of the two tests are superficially very similar. Four factors have loadings above 0.20 based on a weighted average of two analyses. In order of importance, these factors are: numerical (0.51), verbal (0.37), mathematics background (0.37), and general reasoning (0.24). It appears that the higher verbal loading of the E form may be explained by the fact that neither the mathematics-background factor nor the general-reasoning factor was isolated in the battery in which the E form appeared. This explanation, if correct, accounts for a large part of the apparent factorial difference between the two forms of the test.

(5) Mathematics A (CI702GX1)[g].—This form is a revision of CI702F, containing 57 items. Prime objectives were (1) to avoid all arithmetic-reasoning content, (2) to include more items in higher mathematics and thereby broaden the base of the test, (3) to increase the difficulty in order to discriminate better among the more capable, and (4) to include items that would be most valid for navigator selection. Internal-consistency item analysis against total scores on Mathematics A, CI702F, yielded a mean phi value of 0.43. Although this figure is lower than the mean phis for previous forms, 35 items have phi values of 0.43 and above and yield a mean of 0.50. Owing to the fact that the use of general

[e] Developed at Psychological Research Unit No. 3. Chief contributors: Maj. Merrill F. Roff, Capt. Harry Rosenberg, Lt. Robert Utter.

[f] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, S/Sgt. Benjamin Fruchter, Sgt. Betty J. Salk.

[g] Developed at Psychological Research Unit No. 3. Chief contributor: Sgt. Betty J. Salk.

mathematics in the classification battery was discontinued, no further use was made of this form of the test.

COMPUTATION TESTS

Numerical Operations, CI701B[a]

This test was constructed to satisfy the need for measurement of performance in the simple arithmetical processes. Presumably, proficiency in these simple operations should have important bearing upon success in other tasks where such operations are involved. Repeated observations of the tasks of bombardier and navigator, the former in particular, led to a growing conviction that the mathematics most significant is computational. This is in view of the liberal aid supplied to the students in the form of tables and other accessories and of rule-of-thumb methods taught in ground schools.

Description. (1) *Internal characteristics.*—This test involves only the four fundamental arithmetical operations. The problems are printed on an expendable IBM answer sheet. The front of the sheet contains 100 addition and multiplication problems to which answers are given. Each answer is followed by two spaces for marking. If the answer is correct, the "C" space is to be blackened; and if the answer is wrong, the "W" space is to be blackened. The use of response "R" for right was avoided to prevent confusion with the other common opposition, right versus left. Answers to the first three items are already marked on the answer sheet to illustrate the method of answering. The back of the test sheet contains 80 subtraction and division problems with 5-alternative, multiple-choice responses. The examinee is to blacken the space for the correct answer. Answers to the first two of these problems are premarked to illustrate the method of answering. Because of the extremely low absolute difficulty of the problems, the time limits were made short, thus making the test highly speeded. The following problems are typical of the content of this test.

Front

		C	W
11+19+22=52			
Multiply:	139 × 7 = 973	W C	
Add:	12 + 19 + 28 = 59	C W	

Back

Subtract: 63—38: 25 21 29 32 26

Divide: 233÷7: 39 37 $33\frac{2}{7}$ $37\frac{2}{7}$ 35

Subtract: 93 − 58 — Answer: 45, 35, 34, 46, 43

(2) *Administration.*—Instructions are printed upside down with respect to the test problems so strict control can be maintained on working

[a] This test embraced minor revisions of Form A which had been constructed by Cooperative Test Service.

time. The examinees are permitted to go on to the back of the sheet if they finish the front before time is called. They are also allowed to go back to any part of the test to check or correct their answers if time is available. The time allowance is 5 minutes each for the front and back of the test sheet.

(3) *Scoring.*—The test was first scored R−3W, and later (R−3W)/2 in order to obtain a smaller range of scores.

Statistical results.—Extended use of this test made possible the accumulation of a large amount of statistical data. Only samples of these data are given.

(1) *Distribution of scores.*—Administered to unclassified aviation students, the test yielded the typical distribution constants given in table 6.5.

TABLE 6.5.—*Distribution Statistics for Numerical Operations, CI701B, using the scoring formula (R - 3W)/2*

Group	Part	N	M	SD
Unclassified aviation students[1]	Front	1,520	16.9	5.8
Unclassified aviation students[2]	Front	2,376	17.5	6.3
West Point cadets, class of 1946	Front	888	22.7	5.6
Unclassified aviation students[3]	Back	1,148	15.4	6.0
Unclassified aviation students[4]	Back	1,143	16.3	6.0
West Point cadets, class of 1946	Back	888	22.3	5.9

[1] Tested in August and September 1942 at Psychological Research Unit No. 3.
[2] Tested in September and October 1942 at Psychological Research Unit No. 2.
[3] Tested in August and September 1942 at Psychological Research Unit No. 2.
[4] Tested in December 1942 at Psychological Research Unit No. 3.

(2) *Optimal scoring formula.*—Studies to determine the scoring formula to maximize validity yielded the results given in table 6.6. For scoring purposes, weights of approximately −3.0 for the front and −2.0 for the back are recommended for the wrongs score when the test is used for the selection of navigators. For bombardiers, weights of approximately −0.5 for the front and 0 for the back are recommended for wrongs score. Table 6.6 gives the data on which these statistics are based.

TABLE 6.6.—*Data pertaining to the derivation of optimal weights for wrongs score of Numerical Operations, CI701B*[1]

Sample	N	Part	M_R	M_W	SD_R	SD_W	r_{RC}	r_{WC}	r_{RW}	R_c	$(R+aW)$[1]
Navigators	838	Front	45.50	1.31	11.12	1.36	0.36	−0.12	0.04	0.38	−2.94
Navigators	838	Back	42.20	1.18	10.34	1.30	.47	−.11	.03	.49	−1.98
Bombardiers	978	Front	44.45	1.65	11.95	1.59	.16	.00	....	.16	−.44
Bombardiers	978	Back	38.10	1.37	10.20	1.47	.18	−.61	....	.18	−.08

[1] Symbols used in this table are as follows: R=rights score; W=wrongs score; C=criterion; and a=weight for wrongs score.

TABLE 6.7.—*Estimates of Reliability of Numerical Operations, CI701B*

Type	N	r'_{tt}	r_{tt}
Separately timed halves (front)[1]	1,176	0.48	0.64
Separately timed halves (back)[1]	1,176	.64	.79
Test-retest (back)	712	....	.83
Test-retest (front)	712	....	.75
Front-back	4,774	.68	....

[1] Special administration carried out at Medical and Psychological Examining Unit No. 8.

TABLE 6.8.—*Validities of Numerical Operations, CI701B, for air-crew tasks (graduation-elimination criterion)*

Group	Class	Research unit	Part	N	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
Pilots in primary training	43–F	2	Front	2,376	0.63	17.6	17.4	6.3	0.02	.....
Pilots in primary training	43–J	1, 2, 3	Front	3,151	.66	17.8	18.0	6.3	−.03	−0.01
Pilots in primary training	43–E	2	Back	1,148	.76	15.2	16.2	6.0	−.10	.....
Pilots in primary training	43–J	1, 2, 3	Back	3,151	.66	15.4	15.3	6.4	.00	.03
Pilots in B-24 transition training	43–J and 43–K	1, 2, 3	Front	983	.92	18.4	16.5	6.3	.15	.....
Pilots in B-26 transition training	43–J and 43–K	1, 2, 3	Front	380	.82	18.4	17.3	6.0	.10	.....
Pilots in P-40 training	43–J and 43–K	1, 2, 3	Front	2,414	.64	17.9	18.0	5.9	−.01	.....
Pilots in B-24 transition training	43–J and 43–K	1, 2, 3	Back	982	.92	16.3	14.4	6.3	.15	.....
Pilots in B-26 transition training	43–J and 43–K	1, 2, 3	Back	380	.82	16.4	15.2	5.9	.11	.....
Pilots in P-40 training	43–J and 43–K	1, 2, 3	Back	2,416	.64	16.0	15.6	6.1	.04	.....
Navigators	43–10 and 43–11	1, 2	Front	730	.87	23.1	21.4	6.1	.14	.30
Navigators	43–12 and 43–15	1, 2, 3	Front	1,969	.79	22.63	19.44	6.27	.29	.41
Navigators	43–10 and 43–11	1, 2	Back	730	.87	21.0	19.1	6.0	.17	.35
Navigators	43–12 and 43–15	1, 2, 3	Back	1,970	.79	20.62	17.55	5.98	.30	.44
Bombardiers[2]	43–8 and 43–11	1, 2, 3	Front	1,829	.79	20.4	19.4	6.3	.10	.....
Bombardiers[3]	43–14 and 43–18	1, 2, 3	Front	455	.84	19.7	17.4	6.2	.20	.22
Bombardiers[4]	43–14 and 43–18	1, 2, 3	Front	524	.86	21.0	19.7	5.9	.12	.16
Bombardiers[2]	43–8 and 43–11	1, 2, 3	Back	1,829	.79	17.7	16.7	6.2	.10	.....
Bombardiers[3]	43–14 and 43–18	1, 2, 3	Back	455	.84	16.9	14.2	6.1	.25	.27
Bombardiers[4]	43–14 and 43–18	1, 2, 3	Back	524	.86	18.7	17.2	5.8	.14	.19

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] New aviation cadets taking 12-week course (without navigation).
[3] New aviation cadets taking 18-week course (with some navigation).
[4] Eliminated pilots taking 18-week course (with some navigation).

Table 6.9.—*Validity data for Numerical Operations, CI701B, using grades in navigator training as criteria, for a sample of 463 trainees in classes 43-10 through 43-15 (at Hondo Army Air Field)*

Criterion	Part	r[1]	$r_{corr.}$[2]
Grades in dead reckoning (ground school)	Front	0.14	0.28
	Back	.28	.43
Grades in celestial navigation (ground school)	Front	.16	.27
	Back	.22	.35
Grades in dead reckoning (flight)	Front	.14	.22
	Back	.18	.26
Grades in celestial navigation (flight)	Front	.13	.23
	Back	.20	.32
Grades in meteorology	Front	.06	.18
	Back	.10	.24
Military grades	Front	.03	.09
	Back	.05	.11
Final composite grades	Front	.17	.31
	Back	.28	.43

[1] Product-moment correlations.
[2] Assumed unrestricted stanine standard deviation not reported.

Table 6.10.—*Validities of Numerical Operations, CI701B, for certain technical specialties*

Specialty	Criterion	Part	N	r
Air mechanic armorer[1]	Average grades	Front	232	0.18
	Average grades	Front	376	.18
Radio operator-mechanic[1]	Pass-fail	Front	235	.44
	Average grades	Front	153	.09
Air mechanic armorer[1]	Average grades	Back	232	.16
	Average grades	Back	376	.04
Radio operator-mechanic[1]	Pass-fail	Back	235	.28
	Average grades	Back	153	.23
Flexible gunnery[2]	Air-to-air	Front	173	.05
Flexible gunnery[3]	Final Examination	Front	173	−.09
Flexible gunnery[3]	Composite ground range	Front	194	−.01
Flexible gunnery[3]	Jam handy trainer	Front	194	−.01
Flexible gunnery[2]	Air-to-air	Back	173	−.03
Flexible gunnery[3]	Final examination	Back	173	−.08

[1] Tested at Psychological Research Units Nos. 1, 2, and 3.
[2] In class 43-48, tested at Psychological Research Units Nos. 1, 2, and 3.
[3] In class 43-45, tested at Psychological Research Units Nos. 1, 2, and 3.

(3) *Reliability.*—Two methods of estimating reliability produced approximately the same results, as indicated by the data in table 6.7. Retesting of the sample of 712 in table 6.7 was done after approximately 30 days' time. Although front-back correlation is not, strictly speaking, an estimate of a reliability, the true reliability of the parts is probably no less than the correlation between them.

(4) *Factorial composition.*—This test is one of the few relatively pure tests. Little or no significant variance appears in any factor other than the one so characteristic of this test and of other mathematical and numerical tests—the numerical factor. In two analyses, in which total score on the test was used for determining intercorrelations, a weighted average of loadings on the factor is 0.66. In these analyses several other factors have slight loadings, but the communality is relatively low (0.58) for the test. In two other analyses, separate front and back scores were used as the basis of intercorrelations. In these, weighted averages of the factor loadings are 0.80 for the front section and 0.82 for the back. Even smaller

amounts of variance are accounted for by other factors, a fact indicated by the communalities of the parts, 0.68 for the front and 0.71 for the back. From these indications, it appears clear that this test is a much purer measure of a single factor than is commonly achieved.

(5) *Test validity.*—In view of the extended use of this test for classification, abundant data are available on its validity. Tables 6.8 to 6.10 give typical validation results for air crew and technical specialties respectively.

Evaluation of the test.—As a relatively pure measure of the numerical factor, this test appears to be the best. As is true for most other tests that have high loadings with that factor, the usefulness of the test is probably restricted to predicting success in pursuits that require rapid use and manipulation of numerical symbols. The actual importance of this function to a task can be ascertained by correlations with a test such as this.

Variations of the test.—Other forms of this test differ in minor respects only from the one just described, as indicated in the following paragraphs.

(1) Numerical Operations, CI701A—Form S.—This form was designed by the Cooperative Test Service and is very similar in all respects to CI701B already described. It was used for classification purposes for a short time prior to the development of form CI701B.

(2) Numerical Operations, CI701BX1[10].—This is an experimental test, developed to measure the numerical factor. It contains 17 addition, 16 multiplication, 16 subtraction, and 16 division problems, plus 8 problems involving more than one process, e. g. $\frac{6{,}125 \times 8}{30 \times 15} = \underline{\qquad\qquad}$.

Multiple-choice answers are listed for all problems in this test, Scores in this form correlate 0.71 with form CI701B front and 0.79 with form CI701B back (N=298 in both cases).

Numerical Approximation, CI706A [11]

This test is designed to measure the student's ability to estimate quickly the accuracy of results of fairly simple arithmetic operations. It differs from numerical operations in that emphasis is placed upon estimation rather than upon computation. Bombardiers and navigators frequently must make computations under pressure and in limited time. It is thus important that they be able to check their work quickly. Gross errors—the most serious ones—are usually detectable because of the unreasonableness of the results. For example, misplaced decimal points and similar errors should be apparent to one who sees the problem as a whole and is able to estimate within reasonable limits the results of arithmetic operations. The cues to discrepancies are many—numbers of digits, size of first and last digits, position of decimal points, and the like.

[10] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. David H. Jenkins, Sgt. Betty J. Salk.
[11] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, Lt. Linn Hutchinson.

Description.—The items of this test require more complex computations than those found in numerical operations and thus simulate more nearly the type of problems familiar to navigators and bombardiers. To minimize the sheer numerical-operation component, directions encourage the examinee to estimate the answers roughly, not taking time to compute them. The time limit also is set so short that those who stop to compute the exact answers inevitably fail to complete enough items to obtain a good score.

(1) *Internal characteristics.*—The test consists of 15 scored items. The processes involved include: (1) Addition, (2) subtraction, (3) multiplication, (4) division, (5) proportions, and (6) roots and powers. In over half of the problems more than one process is involved, as in the following examples:

1,000:2 = 9,950:——

A. 1.89.
B. 4.975.
C. 9.95.
D. 19.9.
E. 49.75.

8,000 × (.96288−.94208) × 1 = ——

A. 20.
B. 80.
C. 160.
D. 208.
E. 344.

(2) *Administration.*—Two sample problems are given in the directions, and the procedures in their solution are explained. Emphasis is placed upon the necessity for speed and the desirability of estimating results rather than computing them exactly. Testing time for the 15 items is 10 minutes.

(3) *Scoring.*—The test is scored with the formula 2R−W/2.

Statistical results.—Although this test appeared for only a short time in the classification battery, considerable statistical data were obtained.

(1) *Distribution statistics.*—Based on administration (at Psychological Research Unit No. 2, in August and September 1942) to a typical sample of 1,520 unclassified aviation students, the test yielded a mean score of 10.9 and standard deviation of 5.6. The distribution was approximately symmetrical.

(2) *Test reliability.*—An odd-even estimate of reliability, on the basis of 200 cases, yielded a corrected coefficient of 0.61. Since the test is speeded, this is an overestimate. The presence of six apparently different types of items and the extreme shortness of the test (15 items) may separately or jointly account for this relatively low figure.

(3) *Factorial composition.*—This test was not included, as such, in any factor analysis, so no information is available regarding its factor content.

703320—47—7

It is probable, however, that the test would have a relatively high loading in the numerical factor best identified with the Numerical Operations test. In the classification battery the Numerical Approximation test was combined with Arithmetic Reasoning to yield a single score for Mathematics B. The higher loading in the numerical factor (0.59) for the Mathematics B test that contained Numerical Approximation than for the Arithmetic Reasoning test alone (0.53) tends to confirm this belief.

(4) *Test validity.*—Validity data obtained for this test are given in tables 6.11 and 6.12.

TABLE 6.11.—*Validity of Numerical Approximation, CI706A, for primary pilot training (graduation-elimination criterion)*

N	p_g	M_g	M_e	SD_t	r_{bis}
1,520[1]	0.75	11.1	10.3	5.6	0.09
1,148[2]	.76	10.7	10.2	5.4	.06

[1] In class 43-D, tested at Psychological Research Unit No. 2.
[2] In class 43-E, tested at Psychological Research Unit No. 2.

TABLE 6.12.—*Validity of Numerical Approximation, CI706A, for prediction of combat-crew training success*

Group	Criterion	N	r
D-8 Bombardier[1]	Pass-fail	675	0.34
D-8 Bombardier[1]	Academic grades	675	.31
D-8 Bombardier[1]	Average circular error	675	.10
Flexible gunnery	Academic grades	from 88 to 131	.38
Flexible gunnery	Ground-to-ground firing		.04
Flexible gunnery	Air-to-air firing		.02

[1] Classes D8-1 to D8-7.
[2] Class 43-4, at Las Vegas Flexible Gunnery School.

(5) *Item validity.*—Although no actual validity data against navigator success are available for this test, items were correlated with average academic grades in navigation preflight training. Results indicated that the test should be a relatively good predictor of success in at least the academic phases of navigation training. The mean phi value for the 15 items was 0.17, the standard deviation was 0.11, and the range from 0.04 to 0.45.

Evaluation.—This test is probably not significantly different in function from Numerical Operations. The question as to which is the purer measure of the numerical factor and which is the better predictor of success in navigation cannot be answered on the basis of available data. If further research should reveal this test to be a purer measure of the numerical factor than Numerical Operations, its usefulness as a selection instrument would be demonstrated. During the period when this test was used in the classification battery, the score in Arithmetic Reasoning (Mathematics B) included the score in this test. For this reason, no independent data for this test were obtained during that period. Factorial content of the composite score is discussed in the chapter on reasoning.

Variations of the test.—Certain preliminary forms of this test were constructed, differing little in purpose or technique from the form already described.

(1) Numerical Approximation, CI706AX1[u].—This first form of the test consists of 30 multiple-choice items of the type described under form CI706A. The form is moderately homogeneous, yielding a mean internal-consistency phi of 0.42 with a standard deviation of 0.10. The difficulty is approximately optimal, the mean difficulty index, corrected for chance, being 0.51 and the standard deviation, 0.21, both based on the proportion of examinees responding to the items.

(2) Numerical Approximation, CI706AX2[u].—This is a revision of the AX1 form, containing 15 items. It does not differ significantly from the classification form (CI706A).

SUMMARY AND EVALUATION OF MATHEMATICS TESTS

As evidenced by the results discussed in this chapter, mathematical and numerical tests proved to be very valid predictors of navigator success. In view of this fact, most of the weight of prediction for navigators rested upon mathematics tests in the early classification batteries.

In contrast to prediction of success in the other two air-crew positions, a high degree of validity in predicting navigation success was attained by using a very limited number of tests. Important among these were Numerical Operations, Numerical Approximation, and General Mathematics tests. After these tests had been used for some time and factorial, as well as validity, data had been gathered, it became evident that there was considerable overlap among them. Because the Numerical Operations test seemed the purest of the three, and because alone it could carry the full burden of measuring the numerical factor, the other two tests were dropped from the battery. Scores in this test were weighted heavily in classification of navigators but less heavily for bombardiers.

Possibly the most important contributions of the research described in this chapter were the discoveries that (1) the numerical factor in itself is exceptionally valid for navigator selection, and (2) that most mathematical tests derive a large part of their validity from this factor. Other valid factors, but much less prominent, in mathematical tests are general-reasoning, mathematics-background, and verbal factors, all of which have some validity for navigator selection. All of these are better measured by means of nonmathematical tests. The only really unique contribution of mathematics tests, then, is the numerical factor. In the light of this fact, it is evident that greatest economy can be achieved by using the purest possible test of that factor. Of the mathematical tests employed in the classification program, the Numerical Operations test appears to be the most satisfactory from this standpoint.

[u] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, Lt. Linn Hutchinson, Maj. Merrill F. Roff.

[u] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, Lt. Linn Hutchinson.

It may be surprising to some that a supposedly highly intellectual task such as mathematics, as measured by Mathematics A, shows significant variance in the verbal, numerical, mathematics-background, and visualization factors only. In view of the estimated reliability of the mathematics test (0.92 for CI702E) and its communality (0.63), it is true that another factor or factors, as yet unidentified, may account for considerable variance of the test. It is significant to note, however, that the validity of the test for navigator selection (0.42) is entirely accounted for by the known factors. It is apparent, then, that whatever now undefined factors enter into the factorial composition of the test, such factors are unrelated to navigation success.

CHAPTER SEVEN

Reasoning Tests[1]

INTRODUCTION

The original impetus for the development of reasoning tests was provided by early formal and informal job analyses that indicated the importance to navigation of accurate reasoning with words and numbers. The job analysis data presented in chapter 1 (see especially tables 1.3 and 1.4) are sufficient support for the expectation that tests of reasoning, especially arithmetic reasoning, would have moderate to high correlations with navigator criteria.

It was originally thought, too, that successful performance in any aircrew position required, among other traits, the ability to reason rapidly and accurately. It was believed that reasoning was involved in many instances of what the pilot instructors called judgment, particularly where decisions were required. The major emphasis placed upon judgment in pilot training justified efforts to discover what types of reasoning tests might cover aspects of judgment. Later job analysis data, however, did not entirely support this line of thought (see tables 1.2 and 1.6) nor did later test results.

While the worth of arithmetic-reasoning tests for selecting navigators became apparent very quickly, they had very little validity for the pilot criterion. The hypothesis was proposed that this failure of reasoning tests to predict success in pilot training was due to the fact that they were couched in verbal and numerical terms, and that neither verbal nor numerical abilities had any relation to the success or failure of pilot trainees. An intensive effort, therefore, was made to develop and validate nonnumerical and nonverbal reasoning tests. Most of the tests discussed in this chapter were developed in this search for a reasoning test valid for pilot selection.

The informed reader will note that most of the tests are not new in type of content or underlying rationale. This is attributable to the fact that reasoning tests had been subjected to a great deal of investigation in previous decades. It was felt desirable, therefore, to adapt the most appropriate of these for the initial study of the relation of nonverbal and nonnumerical reasoning tests to pilot success.

Reasoning tests that involve numerical and verbal variance will be discussed first. Nonverbal, nonnumerical, reasoning tests will then be discussed, following which will be presented a factor analysis of both types of tests.

[1] Written by Capt. John I. Lacey and Cpl. James F. Tait.

Arithmetic Reasoning, CI206C [2]

The predominance of mathematics in the training and duties of navigators insured the development of some type of mathematics test in the initial phases of the classification and selection program. The first classification battery included a mathematics test which, after several transitional forms, became General Mathematics, CI702E.[3] The early forms of this test included both arithmetic-reasoning problems that could be solved with minimal formal mathematical training and achievement problems requiring the use and understanding of at least high-school mathematics, e. g., logarithms, algebraic manipulations. The test was weighted more heavily with the latter type of problem, thus making it primarily an achievement test. It was thought desirable to construct separate tests, one an achievement test, and the other an arithmetic-reasoning test. Arithmetic Reasoning, CI206C, is the final form of the latter type of test. It was designed to be a more difficult form than its immediate predecessors, CI206A and B, in order to provide better differentiation among superior candidates for training in navigation. The test was included in the Classification Battery of July 1943, and it has been used since that time.

Description.—The test consists of 30 arithmetic-reasoning problems. As examples of the test problems, an easy problem and a difficult problem follow.

If a plane is to fly 132 miles in 45 minutes, what must be its average speed?

A. 146.7 m. p. h.
B. 164 m. p. h.
C. 165 m. p. h.
D. 176 m. p. h.
E. 182 m. p. h.

A plane traveled a certain distance from the base at an average rate of 225 miles per hour. Engine trouble forced it to return at an average rate of 150 miles per hour. It left at 11:35 a.m. and returned at 12:05 p.m. How far away from its base was the plane when it turned back?

A. 30 miles
B. 45 miles
C. 50 miles
D. 75 miles
E. 90 miles

(1) *Internal characteristics.*—The items of the test are arranged roughly in order of increasing difficulty. They are formulated in aviation terms in the interest of face validity. All problems are presented simply and concisely, in an attempt to minimize verbal variance.

(2) *Administration.*—The test is printed as the second half of a booklet, the first half of which is Mathematics, CI702F.[3] The first half is known as Mathematics A, and the second half, as Mathematics B. The

[2] Developed at Psychological Research Unit No. 3. Chief Contributors: Capt. Lloyd G Humphreys, Lt. David H. Jenkins, Jean R. Lyons.
[3] See chapter 6 for a discussion of this test and its predecessors.

two tests are timed separately. Scratch paper is provided to all examinees. The time limit for Arithmetic Reasoning is set at 35 minutes.

(3) *Scoring.*—The scoring formula is 2R—W/2.

Statistical results.—Due to its inclusion in the classification battery, voluminous statistical data are available on this test. Typical data are given below.

(1) *Distribution statistics.*—Typical distribution statistics are given in table 7.1.

TABLE 7.1.—*Distribution constants for Arithmetic Reasoning, C1206C*

Group	N	M	SD
Unclassified aviation students[1]	1,920	10.8	8.9
Unclassified aviation students[2]	1,500	14.9	9.3
Unclassified aviation students[3]	3,000	14.5	9.2
West Point cadets, Class of 1946	883	29.6	11.0

[1] Tested with the November 1943 Classification Battery at Medical and Psychological Examining Units Nos. 4 through 10.
[2] Tested with the November 1943 Classification Battery at Psychological Research Units Nos. 1, 2, and 3.
[3] Tested with the July 1943 Classification Battery at Psychological Research Units Nos. 1, 2, and 3.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi coefficient of 0.50, a standard deviation of 0.11, and a range of values from 0.17 to 0.75. These statistics are based upon the responses of the highest 25 percent and the lowest 25 percent in total score of a group of 480 unclassified aviation students, tested on June 23 and 24, 1943, at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—Two estimates of reliability are given in table 7.2.

TABLE 7.2.—*Reliability data for Arithmetic Reasoning, C1206C, based upon samples of unclassified aviation students*

N	Type	r'_{tt}	r_{tt}
500[1]	Odd-even	0.62	0.77
1,000[2]	Equivalent halves	.72	.84

[1] Tested at Medical and Psychological Examining Unit No. 10 with the November 1943 Battery.
[2] Tested at Medical and Psychological Examining Unit No. 7 from January 30, 1944 to Feb. 14, 1944.

(4) *Difficulty.*—Based upon item analysis of the papers of 1,292 classified pilots, the mean proportion of correct responses, corrected for chance success, is 0.57, with a standard deviation of 0.30 and a range from 0.00 to 0.92.

(5) *Factorial composition.*—The most significant loadings are in the numerical (0.48), general-reasoning (0.47), verbal (0.27), and visualization (0.19) factors. It is important to note that the test has a loading of only 0.12 in the mathematical-background factor. The communality is 0.72, to be compared with the two reliability estimates of 0.84 and 0.77. For a full picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Validation results based on several samples are given in table 7.3.

TABLE 7.3.—*Validity Data for Arithmetic Reasoning, CI206C*

Group	Criterion	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$
Pilots in primary training[1]	Graduation-elimination	4,779	0.88	15.72	14.57	9.02	0.07	[2]0.08
Pilots in primary training[3]	Graduation-elimination	2,346	.74	15.57	14.76	9.04	.05	[2].07
Pilots in primary training[4]	Graduation-elimination	3,146	.84	14.41	13.09	9.04	.08	[5].14
Pilots in primary training[6]	Graduation-elimination	1,823	.80	16.90	15.48	8.92	.09	[2].10
WASPs[7]	Graduation-elimination	104	.61	12.84	7.66	8.46	.38	....
Armorers in training[8]	Average grades	269	....	.....	.....	....	[9].24	....
Officer candidates[10]	Eighth week academic average	343	....	.....	.....	....	[9].40	....

[1] In Class 44–E, tested with the July 1943 Classification Battery at Psychological Research Unit No. 3.
[2] Assuming an unrestricted stanine standard deviation of 2.00.
[3] In Class 44–E, tested with the July 1943 Classification Battery at Psychological Research Unit No. 1.
[4] In Class 44–I, tested at Psychological Research Units Nos. 1, 2, and 3.
[5] Assuming an unrestricted stanine standard deviation of 1.87.
[6] In Class 44–E, tested with the July 1943 Classification Battery at Psychological Research Unit No. 2.
[7] In Class 44–W–8, tested by Medical and Psychological Examining Unit No. 8.
[8] Tested at Medical and Psychological Examining Units Nos. 1 through 10. In Lowry Field armament Classes 34–44A and 35–44A.
[9] Product-moment correlation.
[10] In training at Miami Beach, Class 44–E. Tested at Medical and Psychological Examining Unit No. 5.

Variations.—Two forms directly preceded Arithmetic Reasoning, CI206C. The initial form, Arithmetic Reasoning, CI206A,[4] was administered experimentally to unclassified aviation students and subjected to item analyses of difficulty and of internal consistency. Items showing the highest internal-consistency phi coefficients and the most appropriate difficulties were combined with carefully selected new items to make the first permanent form of the test, CI206B.[4] This form entered the classification battery in August 1942, to be replaced in July 1943 by Form C.

Description.—Form B, like Form C, has 30 arithmetic-reasoning problems.

(1) *Internal characteristics.*—The items of this test are also arranged roughly in order of increasing difficulty and expressed simply and concisely in aviation terms.

(2) *Administration.*—Arithmetic Reasoning, CI206B, like Form C, was administered with an achievement test in the same test booklet. The achievement test was Mathematics, CI702E,[5] and was known as Mathematics A. Mathematics B included not only Arithmetic Reasoning, CI206B, however, but also Numerical Approximations, CI706A.[5] Mathematics B was administered with a time limit of 35 minutes.

(3) *Scoring.*—Two equivalent scoring formulas were used: R—W/4 and 2R—W/2. From August 1942 to December 1942, separate scores were secured for Numerical Approximations and for Arithmetic Reasoning. From December 1942 to July 1943, the score for Mathematics B was the sum of the unweighted component scores in the two tests.

[4] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Milton Burdman, Capt. Lloyd G. Humphreys, and Maj. Merrill F. Roff.
[5] See chapter 6 for a discussion of this test.

Statistical results.—Statistical data are available both for Arithmetic Reasoning, CI206B, alone, and for the combination of it with Numerical Approximations, CI706A.

(1) *Distribution statistics.*—Typical distribution statistics are presented in table 7.4.

TABLE 7.4.—*Distribution constants for Arithmetic Reasoning, CI206B, based upon samples of unclassified aviation students (scored R-W/4)*

N	M	SD
[1]1,520	12.7	5.5
[2]2,376	12.4	5.5

[1] Tested in August and September 1942 at Psychological Research Unit No. 2.
[2] Tested in October 1942 at Psychological Research Unit No. 2.

(2) *Internal consistency.*—Data are available for arithmetic reasoning scores alone. For a sample of 400 unclassified aviation students the mean phi coefficient was 0.42, with a standard deviation of 0.14 and a range from 0.10 to 0.73. These data are based on the highest 25 percent and the lowest 25 percent of the groups in total score.

(3) *Reliability coefficient.*—A sample of 200 unclassified aviation students yielded an odd-even estimate of reliability of 0.80, corrected for length, for the combination of CI206B and CI706A.

(4) *Difficulty.*—For a sample of 400 unclassified aviation students, the test (CI206B alone) yielded a mean proportion of correct responses of 0.52, corrected for chance success, with a standard deviation of 0.21 and a range from 0.05 to 0.89.

(5) *Factorial composition.*—The most significant loadings of Mathematics B (combined scores) are in the numerical (0.57), general-reasoning (0.40), verbal (0.29), and visualization factors (0.22). The communality is 0.68, compared to an estimated reliability of 0.80. Form CI206B alone was analyzed in matrices in which the numerical factor was not defined. The comparable loadings are: general reasoning, 0.57, ver-

TABLE 7.5.—*Validation data for Arithmetic Reasoning, CI206B, based on the graduation-elimination criterion*

Group	N_t	p_s	M_s	M_e	SD_t	r_{bis}	$_c r_{bis}$
Navigation students[1]	[3]1,970	0.79	47.54	40.54	12.82	0.32	0.48
Navigation students[3]	[2]731	....	.....	.....	.....	.32	.50
Pilots in primary training[4]	1,520	.75	13.1	11.5	5.5	.17	....
Pilots in primary training[5]	1,148	.76	12.3	12.2	5.3	.01	....
Pilots in basic training[6] ..	1,429	....	13.0	12.2	5.3	.08	....
Bombardier students[7]	552	.84	10.1	9.2	4.9	.11	....
Bombardier students[8]	496	.82	13.4	11.9	5.3	.15	....

[1] In Classes 43-12, 43-13, 43-14, and 43-15. Tested at Psychological Research Units Nos. 2 and 3.
[2] Using combined scores in Arithmetic Reasoning, CI206B, and Numerical Approximation, CI706A.
[3] In Classes 43-10 and 43-11. Tested at Psychological Research Units Nos. 1 and 2.
[4] In Class 43-D. Tested Aug. 6 to Sept. 8, 1942, at Psychological Research Unit No. 2.
[5] In Class 43-E. Tested Aug. 6 to Sept. 8, 1942, at Psychological Research Unit No. 2.
[6] In Class 43-F. Tested at Psychological Research Unit No. 2.
[7] In Class 43-5-7. Tested at Psychological Research Unit No. 1.
[8] In Class 43-5-7. Tested at Psychological Research Unit No. 2.

bal, 0.29, and visualization, 0.10. The communality is 0.51. For a full picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Validation data are presented in tables 7.5 and 7.6.

TABLE 7.6.—*Validation data for combined scores in Arithmetic Reasoning, CI206B, and Numerical Approximation, CI706A, against seven navigation grades for a sample of 463 navigation trainees*[1]

Criterion	r[2]	$r_{corr.}$[3]
Grades in Dead Reckoning (ground school)	0.37	0.52
Grades in Celestial Navigation (ground school)	.29	.42
Grades in Dead Reckoning (flight)	.23	.31
Grades in Celestial Navigation (flight)	.26	.38
Grades in Meteorology	.16	.32
Military Grades	.12	.19
Final Composite Grades	.38	.53

[1] In Hondo Classes 43-10 through 43-15. Tested at Psychological Research Units Nos. 1, 2, and 3.
[2] Product-moment correlations.
[3] Assumed unrestricted stanine standard deviation not reported.

(7) *Item validity.*—Based on a sample of 1,392 classified pilots, and using graduation-elimination from primary training as the criterion (1,033 graduates), the mean validity phi coefficient was 0.10, with a standard deviation of 0.05 and a range from 0.00 to 0.27.

Evaluation of Arithmetic Reasoning, CI206B and C.—Arithmetic reasoning tests are among the most valid predictors of success in navigation training. They are exceeded in that function in the classification battery only by the Dial and Table Reading tests (see ch. 16.). This validity is due primarily to the tests' loadings in the numerical, reasoning, and verbal factors, and to a small degree to the visualization loading. These factors account for the following percentages of the variance of form CI206C: 23 percent, 22 percent, 7 percent, and 4 percent respectively.

What small pilot validity the tests have is due to visualization and spatial loadings, the other factors having no validity for pilots.

The data on the validity for armorers in training and for officer candidates show, as might be expected, that the test is very useful for evaluating general academic aptitude.

It is interesting to compare the factorial composition of Mathematics, CI702F, and Arithmetic Reasoning, CI206C. The reader will remember that the former was designed to measure mathematical achievement, and the latter, quantitative reasoning ability. Mathematics, CI702F, has a loading of 0.24 on the reasoning factor, whereas Arithmetic Reasoning has an average loading of 0.47. The achievement test has a loading of 0.37 on the mathematical-background factor; the Arithmetic Reasoning test has a loading of only 0.12, which might be a chance deviation from zero The intentions underlying the development of the two tests, therefore, were realized fairly successfully.

It should be noted that far better tests of the numerical, verbal, and visualization factors exist, but that the arithmetic-reasoning tests best define the general-reasoning factor, albeit with moderate loadings. It is hoped that a pure test of this factor will be found. When it is, arithmetic-reasoning tests will lose their importance.

Number Series, CI215AX1 *

The development of a number-series test was undertaken primarily for the purpose of analyzing the area of nonverbal reasoning. There was no expectation that the test would add to the combined validity, especially for navigators, of already existent numerical and reasoning tests. The test, based on the well-known number-series completion concept, promised low verbal content and a high loading in a reasoning factor.

Description.—Each problem in the test consists of an incomplete number series. It is the task of the examinee to determine by what rule of progression the series was constructed, and then to fill the gaps left in the progression with the missing numbers. Since two numbers are omitted in each progression, a problem contains two separately scored responses. Since the difficulty of a problem is largely dependent upon the determination of a rule of progression, and not upon the simple arithmetic involved, the examinee usually answers the items of a problem as a pair, correctly or incorrectly.

(1) *Internal characteristics.*—The test is divided into 2 parts, each part containing 19 problems (38 scored responses). Part I also contains two unscored sample problems. These sample problems are reproduced below, with accompanying text from the directions.

Sample Problems 1 and 2:

4 6 ____ 10 12 ____

(1)		(2)	
A.	4	A.	10
B.	5	B.	14
C.	8	C.	15
D.	12	D.	16
E.	20	E.	18

The series above consists of numbers which increase by twos.

Therefore, the answer to problem 1 is 8, and the answer to problem 2 is 14.

Sample Problems 3 and 4:

29 22 16 11 7 ____ ____

(3)		(4)	
A.	0	A.	1
B.	1	B.	2
C.	3	C.	3
D.	4	D.	4
E.	10	E.	5

These numbers decrease by an amount which each time is decreased by one. That is, 22 is 7 less than 29, 16 is 6 less than 22, 11 is 5 less than 16, and 7 is 4 less than 11. Now continuing the series, 4 is 3 less than 7, and 2 is 2 less than 4. Therefore, the answer to sample problem 3 is 4, and you should have blackened the space under D on your answer sheet. The answer to problem 4 is 2, and you should have blackened the space under B on your answer sheet.

* Developed at Psychological Research Unit No. 3. Chief contributors: Lt. David H. Jenkins and Jean R. Lyons.

(2) *Administration.*—A brief statement of the examinee's task and the two sample problems are printed on the cover of the test booklet. This, the explanation and solution of these problems, and a paragraph cautioning the examinees to avoid sheer guessing constitute the formal administrative directions for the test. The two parts of the test are given and timed separately. The time limit for part I is 14 minutes; for part II, 18 minutes. The difference in time allotted to the two sections allows for the increasing difficulty of the problems. Directions for the test can be administered in 4 minutes, bringing the total testing time to 36 minutes.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—Based, for the most part, on single samples of moderate sizes, the data for this test are relatively complete but not sufficiently extensive to be conclusive.

(1) *Distribution statistics.*—A sample of 194 classified pilots (class 44–A) yielded a mean score of 47.9, with a standard deviation of 11.8. The distribution curve is negatively skewed and somewhat flatter than normal.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.30, with a standard deviation of 0.17 and a range of values from 0.00 to 0.84. These statistics are based upon an analysis of the responses of the highest 25 percent and the lowest 25 percent in total score of a group of 480 unclassified aviation students, tested at Psychological Research Unit No. 3 in April 1942 and May 1943.

(3) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.75, corrected for length, was obtained. This figure is based on a sample of 204 unclassified aviation students.

(4) *Difficulty.*—Based upon item analysis of the responses of the 480 unclassified aviation students mentioned above, the test yielded a mean proportion of correct responses of 0.76, corrected for chance, with a range from 0.20 to 0.99 and a standard deviation of 0.21. For part I the mean is 0.80, with a range from 0.25 to 0.99 and a standard deviation of 0.20. For part II the mean is 0.70, range 0.20 to 0.99, standard deviation 0.24.

(5) *Factorial composition.*—Significant loadings appear only in the numerical (0.47) and general-reasoning (0.36) factors. The communality is 0.47. For a full picture of the factorial composition of this test, see Appendix B.

(6) *Test validity.*—Validation results are presented in table 7.7.

TABLE 7.7.—*Validation data for Number Series, CI215AX1*

Group	Criterion	N_t	p_s	M_s	M_e	SD_t	r_{bis}	$_cr_{bis}$
Pilots in primary training[1]	Graduation-elimination	194	0.86	48.70	42.95	11.75	0.27	[2]0.31
Navigation students	Flight mission grades	200	....	.....	.....	.....	[3].15	.....
Navigation students[4]	Ground mission grades	200	....	.....	.....	.....	[3].13	.....
Navigation students[4]	Weighted average grades	200	....	.....	.....	.....	[3].25	.....

[1] In class 44A. Tested at Psychological Research Unit No. 3.
[2] Assuming an unrestricted stanine standard deviation of 2.00.
[3] Product-moment correlation.
[4] Same sample as above.

Evaluation.—Only 47 percent of the total variance of the test is accounted for by the common factors extracted in an analysis of the nonverbal reasoning battery, to be described later in this chapter. Since the test has a fairly high reliability (0.75), there remains a substantial amount of undefined nonerror variance. Future research should attempt to define this unknown variance.

Nineteen percent of the total test variance is attributable to the numerical factor and 13 percent to the general-reasoning factor. Much better tests of these factors exist.

The navigator validity of the test is moderate and is due primarily to the test's loadings in the numerical and reasoning factors. The validity to be expected from these two factors alone would be close to 0.30, which is notably higher than the obtained validities.

The pilot validity of 0.31 was found for a small sample, and, judging from the factorial composition of the test, is in considerable error. A pilot validity of 0.12 was found for a comparable form (see below) on a much larger sample of 2,115 cases. The weighted average of these validities is 0.13. The pilot validity expected for this test, based upon factor estimates, is 0.04, leaving much obtained validity to be accounted for by unknown factor variance. For this reason the test deserves further analytical study.

Reasoning Test, CI215A

This version of the Number Series Test differs from the CI215AX1 form in directions and in the number and specific content of the problems.

Evidence from factor analysis indicates that the numerical and reasoning factors, which chiefly characterize the Number Series Test, are not related to pilot success. It was thought that a modified form of the test might have a sufficiently high correlation with the pilot stanine and a sufficiently low one with the pilot primary graduation-elimination criterion to justify its inclusion in the classification battery with a negative weight assigned for pilots.

New directions, accordingly, were written to give the test a "pilot slant," the purpose being to prevent men with a strong preference for pilot training from slighting the test.

Description.—The cover of the test booklet, formerly carrying test directions, now portrays a full-page picture of two United States pursuit planes and a burning enemy craft. The directions, formerly a terse outline of the test-task, were increased by 170 words devoted to the relationship of the test to pilot and other air-crew duties. The number of test problems was reduced to 25 (50 scored responses). Testing time, including 3 minutes for administration, totals 23 minutes. The scoring formula is R−W/4.

Statistical results.—The available data are restricted to distribution statistics, item difficulty, and validity.

(1) *Distribution statistics.*—A sample of 1,390 classified pilots (class 44G, tested in January 1944 at Psychological Research Unit No. 3) yielded a mean score of 21.6, a standard deviation of 5.80.

(2) *Difficulty.*—Based on a sample of 728 classified pilots, the test yielded a mean proportion of correct responses of 0.52, corrected for chance success, with a standard deviation of 0.31 and a range from 0.01 to 0.96.

(3) *Test validity.*—For a sample of 2,115 classified pilots (class 44G, tested at Psychological Research Unit No. 3), using graduation-elimination from primary training as the criterion, the uncorrected biserial *r* was 0.10; corrected for restriction of range, the validity was 0.12. Of this sample, 89 percent was graduates. The mean score of graduates was 20.83, of eliminees 19.68, and the over-all standard deviation was 5.60. For this same sample, the correlation with pilot stanine was 0.23, corrected for restriction of range.

(4) *Item validity.*—For a sample of 600 graduates and 128 eliminees from primary training (class 44G), the mean phi coefficient was 0.04, the standard deviation 0.06, and the range from −0.08 to 0.18.

Evaluation.—The validity of 0.12, compared with the correlation of 0.23 with the pilot stanine, precludes the use of this test as a suppression variable for pilot selection. This test should be factorially similar to Number Series, CI215AX1.

Logical Sequence (Numerical Sequence), CI217A

This test was developed at Tuskegee Army Air Field for possible use in the classification of Negro air crew. It is in completion form rather than multiple-choice form. Initial informal reports of exceedingly high validities against a pilot criterion for Negro aviation students were made. Since multiple-choice reasoning tests were known to have so little validity for pilot selection, the test was forwarded by Headquarters AAF Training Command to Psychological Research Unit No. 3 for study.

Some items were added to the test, and with Pattern Sequence, CI217B, it was administered in an intercorrelational study, designed to reveal whether utilizing free-response rather than multiple-choice forms of a test changes factorial composition.

Description.—The test is a typical number-series test, but it varies in form of presentation from Number Series, CI215AX1, described above.

(1) *Internal characteristics.*—The number series are punctuated in a manner that assists the examinee in understanding the internal relationships of the digits. For example, one problem reads as follows: 13–10; 11–7; 9–4. The examinee supplies the next two numbers.

Unlike most printed tests developed by the Aviation Psychology Program, problems of the Numerical Sequence test are not answered by the selection of one or more prepared alternatives. In place of the standard IBM answer sheet, a special blank is provided, and the answer to each item must be written by the examinee.

The test is made up of 1 sample problem, 1 practice problem and 40 scored items. The scored items are divided equally between two separately-timed parts of the test. The number of digits in each problem ranges from 6 through 10, and the items are arranged in approximate order of increasing difficulty.

(2) *Administration.*—Administrative directions for this test are short but adequate. The test is explained with the assistance of one sample problem and one practice problem.

The total testing time is 13 minutes: directions, 3 minutes; part I, 5 minutes; part II, 5 minutes.

(3) *Scoring.*—For purposes of analysis both the number of correct and the number of incorrect responses are recorded for this test.

Statistical results. (1) *Test validity.*—Test validity data are available for Negro trainees. They do not support the initial claim of high validity for this test. For a group of 468 graduates and 217 eliminees, using the primary graduation-elimination criterion, the uncorrected biserial *r* was 0.04. The mean score of graduates was 7.22, of eliminees 6.98, and the over-all standard deviation was 3.45.

(2) *Intercorrelations.*—Some selected intercorrelations are shown in table 7.8. These data allow a comparison between two tests (Numerical Sequence, CI217A, and Reasoning, CI215A) that presumably would be very similar factorially, except for possible differences attributable to the different modes of presentation (free-response *v.* multiple-choice).

TABLE 7.8.—*Product-moment correlations of Numerical Sequence, CI217A, and Reasoning, CI215A, with selected tests (N=353 unclassified aviation students)*[1]

Test	Correlations with		
	Reasoning	Numerical Sequence	
	(R−W/4)	R	W
Numerical Operations (Front), CI701B	0.36	0.50	−0.20
Numerical Operations (Back), CI701B	.38	.48	−.14
Dial and Table Reading, CP621-622A	.44	.53	−.22
Speed of Identification, CP610A	.16	.22	.07
Spatial Orientation I, CP501B	.23	.26	−.02
Spatial Orientation II, CP503B	.18	.07	−.10
Arithmetic Reasoning, CI206C	.43	.50	−.26
Reading Comprehension, CI614H	.37	.40	−.17
Reasoning, CI215A	....	.54	−.14

[1] Tested in October 1944 at Medical and Psychological Examining Unit No 8.

Evaluation.—The test's pilot validity was overestimated in early reports. The validity coefficient of 0.04 reported for a fairly large sample is in accord with expectations for a numerical and reasoning test.

The data in table 7.8 reveal some interesting differences between the multiple-choice, number-series test (Reasoning, CI215A) and the completion form (Numerical Sequence). The higher correlations of Numerical Sequence with Numerical Operations, front and back, and with Dial and Table Reading,[7] leave little room for doubt that it has a higher loading on

[7] See chapter 28 for a complete description of the factorial composition of these tests.

the numerical factor than does Reasoning, CI215A. Whether this is entirely due to the difference between multiple-choice and free-response forms is a moot question, since there are other minor differences between the two tests. This explanation, however, seems reasonable.

The slightly higher correlations of Numerical Sequence with Speed of Identification and Spatial Orientation I,[7] again, suggest a higher loading of the free-response form on the perceptual-speed factor. The lower correlation of the test with Spatial Orientation II, however, casts doubt upon this conclusion.

The higher correlation of Numerical Sequence with Arithmetic Reasoning could be due either to an increased saturation with the numerical factor or with the general-reasoning factor. The very slight increase in correlation with Reading Comprehension, however, suggests that the latter interpretation is more likely.

The correlation of Numerical Sequence with Reasoning, CI215A, is only 0.54. Unless the former test is quite unreliable, this suggests less communality between the two tests than should be expected.

These data, of course, are more suggestive than they are conclusive.

NONNUMERICAL, NONVERBAL REASONING TESTS

Decoding, CI214AX2 [8]

This is one of the battery of nonverbal, nonnumerical tests of reasoning ability, developed in the hope of finding a reasoning ability that would be valid for pilots. It should be noted that the terms nonverbal and nonnumerical, as applied to this and other tests discussed in this section, do not mean that words and numbers do not enter into the test. They do signify that the test was constructed to minimize numerical and verbal variances.

Description.—The test requires the decoding of short words written in a code of signal flags. The items are arranged in groups. In each group

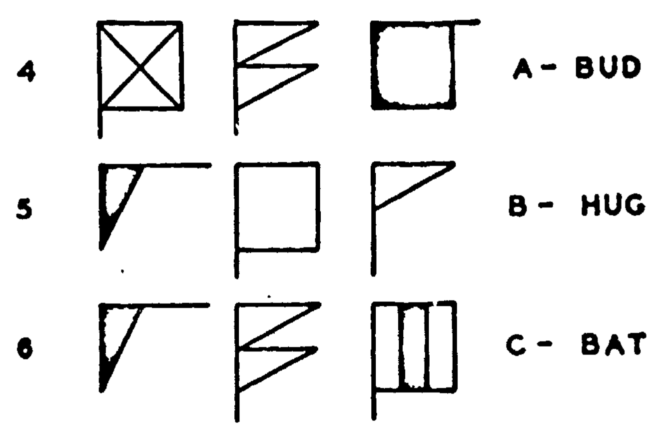


FIGURE 7.1
SAMPLE PROBLEMS OF DECODING TEST,
CI214AX2

[8] Developed at Psychological Research Unit No. 3. Chief contributors: S/Sgt. J. Gordon Eakin, Jean R. Lyons.

three to six rows of individual flag symbols are presented. Each symbol represents an unknown letter of the alphabet. They are arranged three or four to the row, and each row forms a commonplace English word when decoded. After examining the position of repeated flag symbols, it is possible to deduce the word that must correspond to each of the symbol lines. To illustrate the type of problem in the test, sample problem II, used in the directions, is shown in figure 7.1. The accompanying text follows:

Note that the letter b appears twice, each time at the beginning of a word. Since the symbol which appears at the beginning of two code words is a black pennant * * * this pennant must represent b. The letter u also occurs twice in these words, both times in the middle of a word. Thus the code symbol for u is the double white pennant * * * Since the first 2 symbols of item 6 are those which stand for b and u, this item must be bud. The other two items are solved by noting that item 5 begins with b and is therefore bat and that item 4 contains u as a middle letter and, therefore, must be hug * * *

An example, illustrative of the higher difficulty levels of the test, is the last problem in the test, shown in figure 7.2.

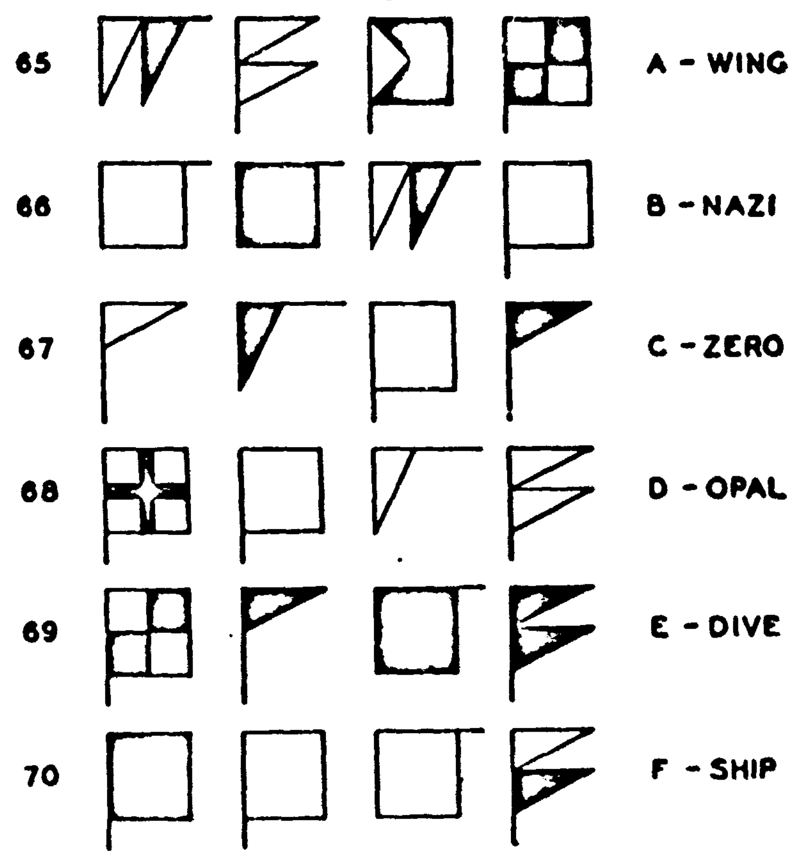


FIGURE 7.2
SAMPLE PROBLEMS OF DECODING TEST, CI214AX2, SHOWING A DIFFICULT PROBLEM

(1) *Internal characteristics.*—There are 11 groups of flag symbols, yielding 64 scored responses.

(2) *Administration.*—Because the task of the examinee is relatively complex, administrative directions for the test are long and detailed.

703320—47—8

The directions consist of a generalization of the test-task and two practice problems with accompanying explanations. The administrator solves the problems with the examinee by following directions printed below each sample problem. Scratch paper is provided to all examinees.

(3) *Scoring.*—The scoring formula is R−W/5.

Statistical results.—This test has appeared in reliability, factor-analysis, and validation studies.

(1) *Distribution statistics.*—Typical distribution statistics obtained on this test are shown in table 7.9.

TABLE 7.9.—*Distribution constants for Decoding Test, CI214AX2, based upon samples of classified pilots*

N	M	SD
[1]563	23.0	10.2
[2]231	26.2	10.7
[3]895	23.4	10.5

[1] In class 44E. Tested at Psychological Research Unit No. 3.
[2] In class 44H. Tested at Psychological Research Unit No. 3.
[3] In classes 44E, 44F, and 44H. Tested at Psychological Research Unit No. 3. Overlaps with two previous samples.

(2) *Reliability coefficients.*—As shown in table 3.1, the administration of separately timed halves of the test yielded an uncorrected reliability of 0.58 for unclassified aviation students and of 0.64 for unclassified aviation students and airplane mechanics, with the coefficient unaffected by the time interval between the administration of the two halves. The corresponding corrected figures are 0.73 and 0.78.

(3) *Factorial composition.*—In one somewhat unsatisfactory analysis (of the November 1943 classification battery, see ch. 28), the test had loadings of 0.32 on the spatial-relations factor and 0.31 on the perceptual-speed factor. The communality was only 0.26. In this analysis, only the perceptual, spatial, social-science background, verbal, mechanical, and mathematical-background factors were defined. A better conception of the factorial composition of a test of decoding may be gained from the discussion of Decoding, CI214AX1, which immediately follows. For a full description of the factorial composition of this test, see Appendix B.

(4) *Test validity.*—Validation results based on several samples are given in table 7.10.

TABLE 7.10.—*Validity data for Decoding Test, CI214AX2*

Group	Criterion	N_t	p_s	M_s	M_u	SD_t	r_{bis}	[1]r_{bis}
Pilots in primary training[2]	Graduation-elimination	231	0.91	26.45	23.75	10.70	0.13	0.16
Pilots in primary training[3]	Graduation-elimination	563	.94	23.00	22.70	10.15	.01	.08
Pilots in primary training[4]	Graduation-elimination	895	.93	23.40	22.85	10.45	.02	.06
Pilots in primary training[5]	Graduation-elimination	443	.87	22.50	19.70	10.60	.14	.20
Navigation students	Flight mission grades	200	....	.....	.....	.....	[6].20	....
Navigation students[7]	Ground mission grades	200	....	.....	.....	.....	[6].24	....
Navigation students[7]	Weighted total grades	200	....	.....	.....	.....	[6].24	....

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44H. Tested at Psychological Research Unit No. 3.
[3] In class 44E. Tested at Psychological Research Unit No. 3.
[4] In classes 44E, 44F, and 44H. Tested at Psychological Research Unit No. 3. Overlaps with two previous samples.
[5] In classes 44I and 44J. Tested at Psychological Research Unit No. 3.
[6] Product-moment correlation.
[7] Same samples as above.

(5) *Item validity.*—Validation of items revealed a mean phi of 0.00, based upon the responses of 600 graduates and 62 eliminees from primary training (class 44H). The standard deviation of phi values was 0.08, and the range was from −0.19 to +0.30.

Evaluation.—Decoding, CI214AX2, shows little pilot validity. The highest corrected validity yielded by any of the numerous samples studied is 0.20 on a total of 483 cases. The weighted average for 1,529 cases is 0.13. In the factor analysis of the battery of nonverbal reasoning tests, Decoding CI214AX1, an earlier form of the test (see discussion immediately following), showed significant loadings in the following factors: general reasoning, reasoning II, reasoning III, perceptual speed, and spatial relations. If a reasoning factor with pilot validity exists, it is not defined by this test in this analysis, for the pilot validity shown by the test can be accounted for by its variance in the perceptual-speed and spatial factors.

The navigator validity (0.24 uncorrected) is expected in view of the test's reasoning, spatial, and perceptual content.

Decoding, CI214AX1

A variation.—This preliminary form of the Decoding test differs somewhat from the final version. It is important primarily because of factorial data available on it.

Description.—This form of the test is divided into 2 comparable parts of 45 scored responses each. Directions and type of items are identical with those in the final form of the test.

(1) *Administration.*—The over-all testing time is 50 minutes; part I takes 25 minutes, part II, 20 minutes, and the directions require 5 minutes.

(2) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—No validation data were compiled on this test in view of anticipated revisions.

(1) *Distribution statistics.*—A sample of 204 unclassified aviation students (tested in May 1943 at Psychological Research Unit No. 3) yielded a mean score of 28.3, a standard deviation of 10.1.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.22, a standard deviation of the phi distribution of 0.13, and a range of values from −0.06 to +0.46. These statistics are based upon analysis of the responses of the highest 25 percent and the lowest 25 percent in total score of the group of 204 unclassified aviation students mentioned above.

(3) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.72, corrected for length, was obtained. This figure is based on the sample of 204 unclassified aviation students.

(4) *Factorial composition.*—The most significant loadings are in the general reasoning (0.36), reasoning III (0.37), perceptual-speed (0.36),

reasoning II (0.30), and spatial-relations (0.19) factors. The communality is 0.54. For a full pict re of the factorial composition of this test see Appendix B.

Evaluation.—Fifty-four percent of the test's total variance is accounted for by the common factors extracted in the analysis to be described later in this chapter, leaving considerable undefined nonerror variance. Significant percentages of the total variance are attributed to the various factors as follows: reasoning III, 14 percent; general reasoning, 13 percent; perceptual, 13 percent; reasoning II, 9 percent, and spatial relations, 4 percent. The remaining variance is spread over other factors in negligible amounts.

The test is important because it helps define the two new factors, reasoning II and reasoning III. (See factor anal sis at end of this chapter.)

Figure Analogies Test, CI212AX1 [a]

This is a variation of the familiar figure-analogies test which has appeared, among other places, in the American Council on Education College Aptitude Test. Generally recognized as a nonverbal reasoning test, this form was developed for inclusion in the analysis of the nonverbal reasoning tests.

Description.—The Figure Analogies Test is designed to measure the ability to formulate correct logical relationships between sets of geometric figures. A test item presents the examinee first with three geometric figures labeled X, Y, and Z, which set the problem, and then with five alternate answers lettered A through E. Figure Y is always a simple variation of figure X. After ascertaining the relationship between the first two figures, the examinee selects from five alternatives the figure that bears the same relation to Z as Y did to X. Sample problem 1, used in the directions, is shown in the top panel of figure 7.3, and a problem from the body of the test in the lower panels. The text for the sample problem follows:

> Your task is to find which one of the five choices at the right goes with figure Z the same way figure Y goes with figure X. Figure X is a circle; figure Y is a similar circle divided into 4 equal parts. The figure that goes with figure Z the same way the divided circle Y, goes with the empty circle, X, is figure A. Of the five choices, figure A is the only one which is divided into four equal parts. So, we can say figure X is to figure Y as figure Z is to figure A. Fill in A after number 1 on your answer sheet.

(1) *Internal characteristics.*—The test is divided into two separately timed parts, each consisting of 30 problems. There are five additional unscored problems that are included in the test's administrative directions as sample and practice problems.

(2) *Administration.*—The test is explained to the examinee with the assistance of two simplified sample problems. He is then allowed 2 minutes to solve three slightly more difficult practice problems and to correct any errors in his work. Fifteen minutes are allowed for completion of each

[a] Developed at Psychological Research Unit No. 3. Chief contributor: Lt. Frank J. Dudek.

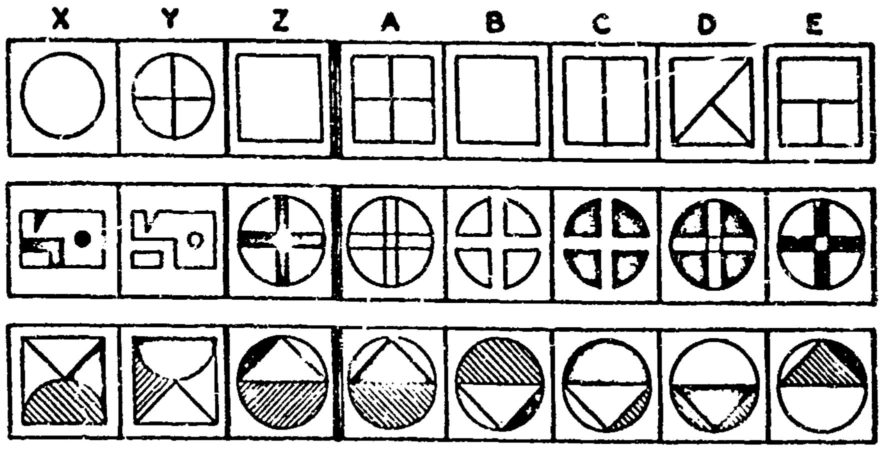


FIGURE 7.3
SAMPLE PROBLEMS OF FIGURE ANALOGIES,
CI212AXI

part. The total testing time, including directions, sample, and practice problems, is 35 minutes.

(3) *Scoring.*—The scoring formula is R − W/4.

Statistical results.—Relatively complete data are available for this test.

(1) *Distribution statistics.*—Typical examples of distribution statistics obtained in this test are given in table 7.11. The distribution curves are slightly negatively skewed.

TABLE 7.11.—*Distribution constants for Figure Analogies, CI212AX1, based upon samples of classified pilots*

N	M	SD
[1]212	34.4	8.2
[2]216	36.1	8.9
[3]496	33.7	8.6

[1] In class 44A. Tested at Psychological Research Unit No. 3.
[2] In class 44B. Tested at Psychological Research Unit No. 3.
[3] In class 44C. Tested at Psychological Research Unit No. 3.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.32, a standard deviation of the phi distribution of 0.12, and a range of values from 0.08 to 0.64. These statistics are based upon analysis of the responses of the highest 25 percent and the lowest 25 percent in total score of a group of 197 unclassified aviation students, tested in March 1943 at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.82, corrected for length, was obtained. This figure is based on a sample of 1,200 unclassified aviation students.

(4) *Difficulty.*—Based upon item analysis of the responses of 197 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.58, corrected for chance success, with a range from 0.00 to 0.96 and a standard deviation of 0.27.

(5) *Factorial composition.*—The most significant loadings are in the reasoning II (0.40), general-reasoning (0.34), integration III (0.34), reasoning III (0.31), visualization (0.28), verbal (0.23), and numerical (0.20) factors. The communality is 0.76. For a full picture of the factorial composition of this test see Appendix B.

(6) *Test validity.*—Validation results based on several samples are given in table 7.12.

TABLE 7.12.—*Validity data for Figure Analogies, CI212AX1*

Group	Criterion	N_t	p_g	M_g	M_e	SD_t	r_{bis}	${}_cr_{bis}$
Pilots in primary training[1]	Graduation-elimination	496	0.93	33.94	31.09	8.64	0.16	.....
Pilots in primary training[2]	Graduation-elimination	712	.92	34.66	32.20	9.48	.13	.....
Pilots in primary training[3]	Graduation-elimination	796	.93	34.12	31.66	10.86	.11	[4]0.19
Pilots in primary training[5]	Graduation-elimination	634	.92	35.13	32.19	8.40	.17	.....
Navigation students	Flight missions	200	....	.....	.....	.....	[6].28	.....
Navigation students[7]	Ground missions	200	....	.....	.....	.....	[6].14	.....
Navigation students[7]	Weighted total	200	....	.....	.....	....	[6].29	.....
Navigation students[8]	Graduation-elimination	1,675	.92	38.39	33.82	7.40	.39	[4].61

[1] In class 44C. Tested at Psychological Research Unit No. 3.
[2] In classes 44B and 44C. Tested at Psychological Research Unit No. 3. Overlaps previous sample.
[3] In classes 44B, 44C, and 44D. Tested at Psychological Research Unit No. 3. Partially overlaps previous sample.
[4] Assuming an unrestricted stanine standard deviation of 2.00.
[5] In classes 44D and 44E. Tested at Psychological Research Unit No. 3. Partially overlaps previous sample.
[6] Product-moment correlation.
[7] Same sample as above.
[8] Tested at Psychological Research Unit No. 1 in June 1944; at Psychological Research Unit No. 2 in May 1944; and at Psychological Research Unit No. 3 in April 1944.

Evaluation.—The weighted averages of the factor loadings of two factor analyses (N=468) account for 76 percent of this test's total variance. The percentages of total variance accounted for are: Reasoning II, 16 percent; general reasoning, 12 percent; integration III, 12 percent; reasoning III, 10 percent; visualization, 8 percent; verbal, 5 percent, and numerical, 4 percent. The remaining variance is spread over other factors in insignificant amounts.

The validity figures are similar to those of other tests in the nonverbal reasoning group. The pilot validity appears to be the result of the combined loadings of several pilot-valid factors, including visualization and perceptual speed. The very much higher navigator validity is to be expected from loadings in the general reasoning, verbal, and numerical factors. The test is also important because it best defines the new factor, reasoning II (see below).

Figure Classification, CI213AX1 [10]

This is a new version of a familiar test. It was developed as a component part of the nonverbal reasoning group.

Description.—As stated in the test's directions, this is a test of the ability to draw comparisons and make generalizations. The task of the

[10] Developed at Psychological Research Unit No. 3. Chief contributor: Lt. Mahlon B. Smith.

examinee is to select from five alternatives the geometric figure that has the characteristic common to each of three figures that set the problem. Practice problem I, used in the directions, is shown in the upper panel of figure 7.4. The explanatory text accompanying this sample problem is:

The three figures to the left of the heavy line, although of different shapes and sizes, are alike in one way. The lines which bound the figure are straight lines. Now examine the five figures labeled A, B, C, D, and E. Find the one figure which is bounded only by straight lines. The only figure which meets this requirement is figure D.

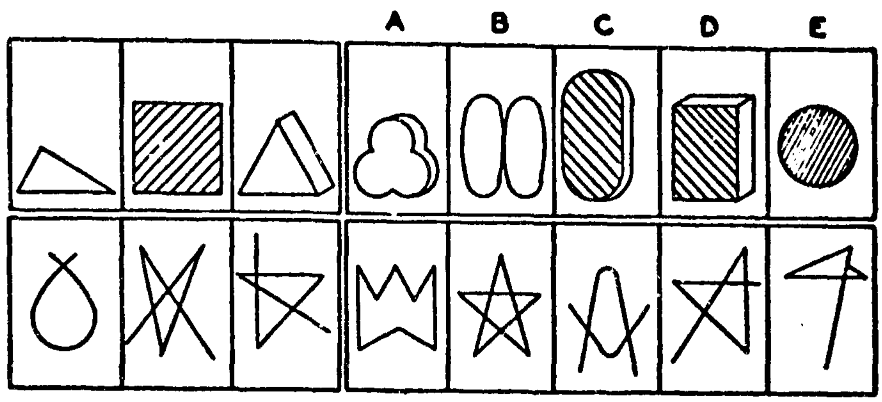


FIGURE 7.4
SAMPLE PROBLEMS OF FIGURE CLASSIFICATION, CI213AX1

The typical test problem requires the detection of exact, but obscure, similarities. At the more difficult levels, the key figures of a test problem appear to bear absolutely no relationship to each other upon initial inspection. Their similarities may exist in such minor characteristics as number of dimensions, type of shading, number of divided areas, type of lines used to enclose the figures, inclusion of certain type and number of angles, etc. An example of one of the more difficult problems is shown in the lower panel of figure 7.4. Figure "D" is the correct answer to this problem. It is the only alternative possessing the characteristic the three key figures have in common; i. e., formation of the figure by use of one continuous line with both ends free.

(1) *Internal characteristics.*—The test is divided into two separately timed parts, each containing 16 items. There are two practice items at the beginning of the test.

(2) *Administration.*—The time limits established for this test are as follows: Directions, 1 minute; part I, 12 minutes; part II, 10½ minutes; over-all testing time, 23½ minutes.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—Extensive data are available for this test.

(1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 7.13. The distribution curves are slightly positively skewed.

TABLE 7.13.—*Distribution constants for Figure Classification, CI213AX1, based upon samples of classified pilots*

N	M	SD
[1]693	13.1	7.0
[2]955	13.0	7.4

[1] In classes 44D and 44E. Tested at Psychological Research Unit No. 3.
[2] In classes 44D, 44E, and 44H. Tested at Psychological Research Unit No. 3. Partially overlaps previous sample.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.49, a standard deviation of the phi distribution of 0.18, and a range of values from 0.10 to 0.77. These statistics are based upon analysis of the responses of the highest 25 percent and the lowest 25 percent in total score of a group of 480 unclassified aviation students, tested in March 1943 at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.78, corrected for length, was obtained. This figure is based on a sample of 440 unclassified aviation students, tested in March 1943 at Psychological Research Unit No. 3.

(4) *Difficulty.*—Based upon item analysis of the responses of 450 unclassified aviation students (tested in March 1943 at Psychological Research Unit No. 3), the test yielded a mean proportion of correct responses of 0.45, corrected for chance success, with a range from 0.05 to 0.99 and a standard deviation of 0.20.

(5) *Factorial composition.*—The most significant loadings are in the integration III (0.38) and reasoning III (0.32) factors. It is important to note that the test has a loading of only 0.03 on the general reasoning factor and of 0.15 on the verbal factor. The communality is only 0.30. For a full picture of the factorial composition of this test see Appendix B.

(6) *Test validity.*—Validation results based on several samples are given in table 7.14.

TABLE 7.14.—*Validity data for Figure Classification CI213AX1*

Group	Criterion	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training[2]	Graduation-elimination	693	0.94	13.08	13.92	6.98	−0.06	−0.03
Pilots in primary training[3]	Graduation-elimination	262	.91	12.72	13.58	8.66	−.05	.02
Pilots in primary training[4]	Graduation-elimination	194	.86	15.82	14.10	6.88	.14	.....
Navigation students	Flight mission grades	200	....	.....	.....	.....	[5].02	.....
Navigation students[6]	Ground mission grades	200	....	.....	.....	.....	[5].13	.....
Navigation students[6]	Weighted total grades	200	....	.....	.....	.....	[5].09	.....

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In classes 44D and 44E. Tested at Psychological Research Unit No. 3.
[3] In class 44H. Tested at Psychological Research Unit No. 3.
[4] In class 44A. Tested at Psychological Research Unit No. 3.
[5] Product-moment correlations.
[6] Same sample as the one preceding.

(7) *Item validity.*—Validation of items revealed a mean phi of 0.02, based upon the responses of 600 graduates and 41 eliminees from training

(in class 44D; tested at Psychological Research Unit No. 3). The standard deviation was 0.07, and the range was from −0.16 to +0.15.

Evaluation.—Although the test has satisfactory reliability, its pilot validity is zero or slightly negative, and its navigator validity is extremely low, if not zero.

The following factors account for the indicated percentages of the test's total variance: verbal, 2 percent; general reasoning, 0 percent; reasoning III, 12 percent; reasoning II, 3 percent; integration III, 15 percent. The zero loading in the navigator-valid general-reasoning factor is mentioned to help interpret the unusually low navigator validity of the test. All factor loadings are critically low, with the possible exception of reasoning III and integration III. These two factors are not known to be valid for any air-crew position.

The common-factor variance represents only 30 percent of the test's total variance. The known factorial content obviously does not present a complete picture of this test. There is little or no pilot or navigator validity to be accounted for by the unknown variance, so this test deserves no further attention in aviation psychology.

Pattern Sequence, CI217B

Like Numerical Sequence, CI217A (see above), this test was developed at Tuskegee Army Air Field for possible use in the classification of Negro aircrew; and it is in completion form, rather than multiple-choice form. It, too, was administered in an intercorrelational study to discover any possible effect of the multiple-choice form upon factorial content.

Description.—Each problem in the test consists of a series of geometric figures, constructed in accordance with a rule of progression. The examinee must determine that rule, and supply the next figure in the series. Thus, the test has some characteristics of the Number Series, Figure Classification, and Figure Analogies tests.

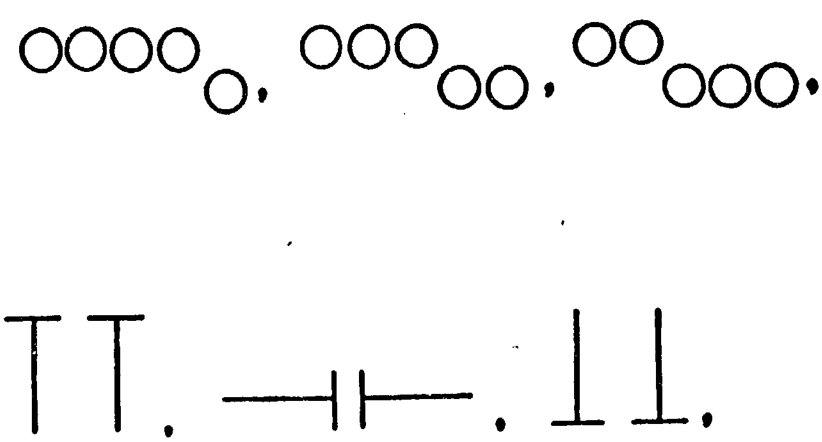


FIGURE 7.5
SAMPLE PROBLEMS OF PATTERN SEQUENCE,
CI217B

(1) *Internal characteristics.*—There are 40 scored items and 1 sample item. The scored items are divided equally between two separately timed parts. Sample problems are shown in figure 7.5.

(2) *Administration.*—One sample and one practice problem compose most of the test's formal directions. Five minutes are allowed for each part.

(3) *Scoring.*—The number of correct and the number of incorrect answers are recorded for this test.

Statistical results. (1) *Test validity.*—Based on a (Negro) sample of 469 graduates and 217 eliminees from pilot primary training, the uncorrected validity was 0.12. The mean score of graduates was 10.44 and of eliminees, 9.76. The over-all standard deviation was 3.36.

(2) *Intercorrelations.*—Selected intercorrelations are shown in table 7.15, comparing Figure Analogies, CI212AX1, and Pattern Sequence.

TABLE 7.15.—*Product-moment correlations of Pattern Sequence, CI217B, and Figure Analogies, CI212AX1, with selected tests (N=353 unclassified students)*[1]

Test	Correlations with		
	Figure Analogies (R−W/4)	Pattern Sequence	
		R	W
Arithmetic Reasoning, CI206C	0.43	0.45	−0.25
Reading Comprehension, CI614H	.41	.42	−.21
Numerical Operations (Front), CI702B	.20	.34	−.15
Numerical Operations (Back), CI702B	.25	.37	−.10
Dial and Table Reading, CP621-622A	.43	.50	−.22
Figure Analogies, CI212AX1	....	.46	−.42

[1] Tested in October 1944 at Medical and Psychological Examining Unit No. 8.

Evaluation.—The similar correlations of Figure Analogies and Pattern Sequence with Arithmetic Reasoning and Reading Comprehension suggest that reasoning variance is not changed by using a nonmultiple-choice form.

The correlations with the Numerical Operations and Dial and Table Reading tests, however, suggest increased variance in the numerical factor of the free-response form. The reader will recall that the same conclusion was drawn in comparing the Number Series and the Number Sequence tests.

As was true for that comparison also, the correlation between the free-response and multiple-choice form is unexpectedly low.

Again it should be stated that these data are merely suggestive, since the tests compared differ in other respects than the use of prepared alternatives.

Spatial Reasoning, CI211BX1 [11]

This test is a revision of the Thurstone marks test and was designed for inclusion in the group of nonverbal reasoning tests.

[11] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter, Jr., and Lt. Linn Hutchinson.

Description.—The test requires the examinee to detect the principle governing the placement of letter symbols in a spatial pattern of dashes and gaps. Figure 7.6 presents sample problem two of the test. The solution for this problem, taken from the test's administrative directions, reads:

The rule for example 2 is, "The number of marks to the left of X increases by one in each row. Y is just to the right of the last gap in each row." X and Y are omitted from the last row, but according to the rule, X should be at the 5th mark (E) and Y should be at the mark following the last gap (I). Blacken spaces E and I after item 2 on your answer sheet.

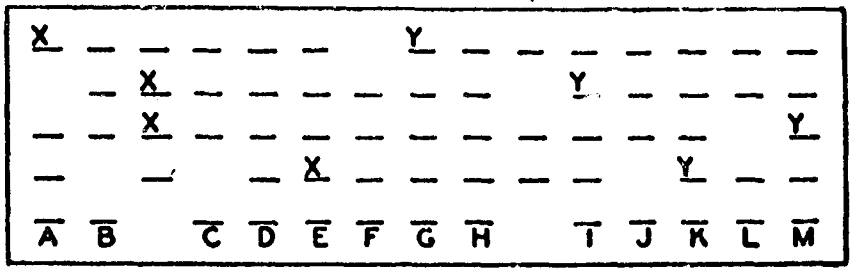


FIGURE 7.6
SAMPLE PROBLEM OF SPATIAL REASONING,
CI211BX1

(1) *Internal characteristics.*—The test is divided into 2 parts, each containing 17 scored items. Some items call for the placement of two symbols, others of three symbols. There is a total of 78 scored responses in the test.

(2) *Administration.*—A general statement of the task involved, a standard paragraph on use of the IBM answer sheet, and three sample problems make up the formal test directions. Total testing time, including 5 minutes for directions and administration, is 50 minutes. The time limit for part I is 25 minutes and for part II, 20 minutes. Three minutes before the end of each period, examinees are informed of the time remaining. A 15-place, IBM answer sheet is used with the test.

(3) *Scoring.*—The scoring formula is R—W/5.

Statistical results.—Owing to the early development of a revised form, only limited data are available on this test.

(1) *Distribution statistics.*—A sample of 224 unclassified aviation students tested at Psychological Research Unit No. 3 in April 1942 yielded a mean score of 47.5, and a standard deviation of 16.1.

(2) *Reliability coefficient.*—A sample of 224 unclassified aviation students tested in March 1942 at Psychological Research Unit No. 3 yielded an alternate-forms correlation of 0.74, which corrects to 0.85.

(3) *Factorial composition.*—The most significant loadings are in the reasoning I (0.45), reasoning III (0.38), integration III (0.38), spatial (0.26), and verbal (0.20) factors. The communality is 0.72. For a full picture of the factorial composition of this test see Appendix B.

(4) *Test validity.*—Validation results based on several samples are given in table 7.16.

TABLE 7.16.—*Validity data for Spatial Reasoning, CI211BX1*

Group	Criterion	N_t	p_g	M_g	M_e	SD_t	r_{bis}
Pilots in primary training[1]	Graduation-elimination	104	0.77	42.60	38.64	20.60	0.11
Navigation students	Flight missions	200	....	.....	.....	.....	[2].21
Navigation students[3]	Ground missions	200	....	.....	.....	.....	[2].26
Navigation students[3]	Weighted total	200	....	.....	.....	.....	[2].27

[1] In class 43I. Tested at Psychological Research Unit No. 3.
[2] Product-moment correlations.
[3] Same sample as that preceding.

Evaluation.—Seventy-two percent of the total variance of the test is accounted for by loadings in common factors extracted in two factor analyses. Significant percentages of variance are attributed to the following factors: reasoning I, 20 percent; reasoning III, 14 percent; integration III, 14 percent; spatial, 7 percent; and verbal, 4 percent. The remaining common-factor variance is spread over seven other factors.

This test has its highest loading (0.45) on the reasoning I factor. Its reliability is satisfactory. As expected for a general-reasoning test, it is valid for navigators, but not for pilots.

Spatial Reasoning, CI211BX2

A variation.—This test is a revision of Spatial Reasoning, CI211BX1, and differs in surface characteristics only. Additional data of value, however, are available on this form of the test.

Description.—Parts I and II of the original test are combined in this form and the total number of scored responses reduced to 70. The x, y, z symbols used to formulate test problems are replaced by numerical digits corresponding to the numbers of the problems. Test directions were rewritten in the interests of clarity, but without major change.

(1) *Administration.*—Over-all testing time was cut from 50 to 30 minutes, with administration time remaining constant.

(2) *Scoring.*—The scoring formula is R−W/5.

Statistical results.—Considerable validity and distribution data were compiled on this form.

(1) *Distribution statistics.*—Typical distribution statistics obtained on this test are given in table 7.17. The distribution curves are somewhat positively skewed and considerably flatter than normal.

TABLE 7.17.—*Distribution constants for Spatial Reasoning, CI211BX2, based upon samples of pilots in primary training*

N	M	SD
[1]104	37.7	20.6
[2]2[illegible]9	46.6	16.5
[3][illegible]5	23.6	14.6

[1] In class 43I. Tested at Psychological Research Unit No. 3.
[2] In class 44A. Tested at Psychological Research Unit No. 3.
[3] In classes 44D and 44E. Tested at Psychological Research Unit No. 3.

(2) *Test validity.*—Validation results based on several samples are given in table 7.18.

TABLE 7.18.—*Validity data for Spatial Reasoning, CI211BX2, using the graduation-elimination criterion*

Group	N_t	p_s	M_s	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training[2]	571	0.94	24.50	22.85	14.65	0.05	0.09
Pilots in primary training[3]	686	.94	23.75	22.10	14.55	.05	.09
Navigation students[4]	1,291	.93	27.98	19.67	14.69	.28	.48

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44D. Tested at Psychological Research Unit No. 3.
[3] In classes 44D and 44E. Tested at Psychological Research Unit No. 3. Overlaps previous sample.
[4] Tested at Psychological Research Unit No. 1 on May 30 and June 1, 1944; at Psychological Research Unit No. 2 in May 1944; at Psychological Research Unit No. 3 in April 1944.

Evaluation.—As indicated by validation data on several samples of substantial size, the test holds little promise as an instrument of pilot selection. Like other tests of the nonverbal reasoning group, the test has substantial navigator validity. The factor pattern for this form of the test should be closely similar to that for the BX1 form.

A FACTOR ANALYSIS OF REASONING TESTS[11]

Despite the large amount of existing research into the psychology of reasoning, but little attention has been paid to the problem of the statistically independent abilities that enter into the solutions of reasoning tasks. The tests described in the previous pages of this chapter all seem to involve thinking through to a solution. Some appear superficially to be tasks of deduction, others of induction. Some require the examinee to adopt and test various hypotheses; some merely set a problem which can be reasoned through to a solution by the application of the rules of mathematics. These and many more aspects of the tests may be delineated by armchair analysis. It is desirable, however, to establish statistical reference points to guide such introspective analysis and, equally important, to secure quantitative evaluations of the contents of the various tests.

The Data

The intercorrelations for this study are based upon a sample of 202 classified pilots who were awaiting entrance to preflight school. At the time of testing (spring of 1943), restriction of range due to disqualification for low aptitude did not constitute a major problem; so the intercorrelations probably are not biased in this sense. The matrix of intercorrelations appears in table 7.19.

Included in the battery of tests are those which were considered, at that time, most promising as nonverbal reasoning tests. These are: Spatial

[11] Accomplished at Psychological Research Unit No. 3. Chief contributors: Capt. Lloyd G. Humphreys and Lt. David H. Jenkins.

TABLE 7.19.—*Correlation matrix for the Nonverbal Reasoning Battery (N=200)*[1]

	Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	Speed of Identification[2]	..	42	22	18	00	05	04	22	10	09	34	26	33	21	37	19	11	29	29	23	28	13
2	Spatial Orientation I	42	..	12	12	07	13	07	24	01	11	32	16	24	11	41	19	09	24	27	17	35	16
3	Vocabulary—Navigator	22	12	..	58	43	24	25	08	14	13	32	32	30	07	28	24	21	37	30	40	27	28
4	Reading Comprehension	18	12	58	..	48	18	46	16	28	23	50	38	40	18	08	35	25	43	47	50	23	33
5	Mathematics B	00	07	43	48	..	53	28	11	32	26	31	19	26	−03	20	41	17	39	30	43	29	53
6	Numerical Operations	05	13	24	18	53	..	04	12	17	20	14	13	03	04	29	26	08	28	10	21	24	41
7	Mechanical Principles	04	07	25	46	28	04	..	24	36	24	33	37	42	20	13	35	14	40	55	48	11	21
8	Complex Coordination	22	24	08	16	11	12	24	..	28	02	33	40	24	13	26	30	12	25	31	27	30	20
9	Air Maneuvers	10	01	14	28	32	17	36	28	..	21	24	37	26	17	23	40	23	33	44	40	20	35
10	Competitive Planning	09	11	13	23	26	20	24	02	21	..	22	09	25	07	07	34	14	36	26	35	29	35
11	Instrument Comprehension I	34	32	32	50	31	14	33	33	24	22	..	48	37	14	21	36	15	43	46	41	39	34
12	Instrument Comprehension II	26	16	32	38	19	13	37	40	37	09	48	..	27	26	27	23	20	40	41	37	25	22
13	Pattern Comprehension	33	24	30	40	26	03	42	24	26	25	37	27	..	28	15	39	19	43	58	58	35	28
14	OP901A, Part III	21	11	07	18	−03	04	20	13	17	07	14	26	28	..	11	13	11	29	34	34	19	12
15	Pursuit	37	41	28	08	20	29	13	26	23	07	21	27	15	11	..	22	13	18	12	15	23	16
16	Spatial Reasoning	19	19	24	35	41	26	35	30	40	34	36	23	39	13	22	..	35	51	50	55	46	40
17	Figure Classification	11	09	21	25	17	08	14	12	23	14	15	20	19	11	13	35	..	34	27	32	16	23
18	Figure Analogies	29	24	37	43	39	28	40	25	33	36	43	40	43	29	18	51	34	..	55	60	44	40
19	Spatial Visualization I	29	27	30	47	30	10	56	31	44	26	46	41	58	34	12	50	27	55	..	68	29	35
20	Spatial Visualization II	23	17	40	50	43	21	48	27	40	35	41	37	58	34	15	55	32	60	68	..	49	42
21	Decoding	28	35	27	23	29	24	11	30	20	29	39	25	35	19	23	46	16	44	29	49	..	30
22	Number Series Completion	13	16	28	33	53	41	21	20	35	35	34	22	28	12	16	40	23	40	35	42	30	..

[1] Decimal points omitted.
[2] For code numbers see table 7.20.

Reasoning, Figure Classification, Figure Analogies, Decoding, Number Series Completion, and the Spatial Visualization tests. All but the Spatial Visualization tests were discussed in this chapter.[13]

In addition to these, seven other experimental tests were included in the battery, although they were not constructed as part of the nonverbal reasoning project. It was hoped, however, that they would clarify the analysis. These tests are: Planning Air Maneuvers and Competitive Planning, discussed in chapter 9; Instrument Comprehension I and II, discussed in chapter 19; Pattern Comprehension, discussed in chapter 12; Pursuit, CP512A, discussed in chapter 16; and Aptitude Test, part III, QP901A. The last test, not treated in this volume, is a Gottschaldt figures test, similar to the test, Camouflaged Outlines, CP821A, discussed in chapter 17.

Eight classification tests were included in the matrix to serve as reference variables. They are: Speed of Identification and Spatial Orientation I, to define the perceptual speed factor; Technical Vocabulary (Navigator) and Reading Comprehension, to define the verbal factor; Mathematics B, for the general reasoning factor; Numerical Operations, for the numerical factor; the SAM Complex Coordination test, to define the spatial relations factor; and the Mechanical Principles test, to define a mechanical factor, but which defined a new factor, as will be seen. All these tests are described fully in this volume, with the exception of the Complex Coordination test. This test is briefly described on p. 122, and fully described in report no. 4 of this series.

Nine factors were extracted and interpreted. The centroid loadings are shown in table 7.20 and the rotated loadings in table 7.21.

The Factors

Rotated factor I is defined by the following tests and loadings:

Test number	Test name	Loading
2	Spatial Orientation I	0.63
1	Speed of Identification	.57
15	Pursuit	.56
21	Decoding	.36
8	Complex Coordination	.27
11	Instrument Comprehension I	.25

This is the familiar perceptual-speed factor, which usually clearly emerges when either Spatial Orientation I or Speed of Identification is in the matrix. In all analyses, one or the other of these tests best defines the factor, with the other test taking second place. The weighted average loadings (see table 28.15) show Speed of Identification to be the slightly better measure of the factor.

The substantial loading on this factor for the Pursuit Test indicates

[13] A description of Spatial Visualization I and II may be found in chapter 12.

TABLE 7.20.—*Centroid loadings and communalities for the Nonverbal Reasoning Battery*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	h^2
1 Speed of Identification, CP610A	41	47	18	−10	−17	−11	09	13	−07	50
2 Spatial Orientation I, CP501B	38	45	[illegible]	−14	09	10	08	09	05	50
3 Technical Vocabulary (Nav), CE505C	52	−14	15	21	−39	09	15	09	18	58
4 Reading Comprehension, CI614G	63	−22	−06	22	−38	24	−02	07	−05	71
5 Mathematics B, CI706A	55	−50	31	17	10	16	05	−11	02	73
6 Numerical Operations, CI701B	38	−25	42	09	18	−16	15	−24	06	53
7 Mechanical Principles, CI903A	53	−08	−38	19	08	24	07	10	−13	56
8 Complex Coordination, CM701A	43	28	−08	17	22	−12	−16	−07	09	40
9 Planning Air Maneuvers, CI408AX3	51	−12	−22	12	29	−17	−05	03	−15	48
10 Competitive Planning, CI409AX2	40	−22	07	−26	07	07	−11	07	−17	34
11 Instrument Comprehension I, CI615A	62	17	04	12	−12	13	−30	−03	−09	56
12 Instrument Comprehension II, CI616B	56	19	−15	36	−10	−16	−14	−14	−06	58
13 Pattern Comprehension, CP803A	62	13	−18	−18	−06	24	14	03	−09	56
14 QP901A—III	33	21	−23	−12	−09	−12	15	−16	−13	31
15 Pursuit (Path Tracing), CP512A	40	23	32	18	16	−17	24	25	12	54
16 Spatial Reasoning, CI211BX1	66	−16	−05	−18	21	−02	−13	13	18	61
17 Figure Classification, CI213AX1	37	−13	−13	−13	−07	−24	−05	18	13	30
18 Figure Analogies, CI212AX1	73	−09	−07	−17	−10	−07	−05	−06	04	60
19 Spatial Visualization I, CI204AX1	74	06	−35	−12	02	17	09	−06	−08	74
20 Spatial Visualization II, CI203AX1	77	−11	−23	−21	−03	11	04	−13	06	74
21 Decoding, CI214AX2	57	15	17	−23	07	06	−16	−14	23	54
22 Number Series Completion, CI215AX1	57	−27	20	−06	12	−06	−06	−11	−15	50

[1] Decimal points omitted.

Table 7.21.—*Rotated factor loadings for the Nonverbal Reasoning Battery*[1]

	Test	I	II	III	IV	V	VI	VII	VIII	IX	h^2
1	Speed of Identification[2]	57	03	−08	06	04	13	31	03	16	48
2	Spatial Orientation I	63	08	16	05	04	00	09	04	−03	45
3	Technical Vocabulary (Navigator)	12	15	15	04	05	66	18	16	01	56
4	Reading Comprehension	−02	−02	33	17	21	65	27	03	11	69
5	Mathematics B	−08	51	55	09	03	34	−04	00	07	70
6	Numerical Operations	05	69	18	05	−09	09	11	07	02	55
7	Mechanical Principles	00	−04	25	29	54	28	01	00	20	56
8	SAM Complex Coordination	27	09	05	52	09	−05	−06	14	10	40
9	Planning Air Maneuvers	00	13	25	41	29	−01	01	−01	33	44
10	Competitive Planning	03	11	41	−05	08	−01	15	01	31	32
11	Instrument Comprehension I	25	04	36	39	03	23	34	00	13	53
12	Instrument Comprehension II	11	15	−01	57	20	23	36	04	01	59
13	Pattern Comprehension	24	−04	33	06	50	18	24	13	13	54
14	CI901A III (Gottschaldt Figures)	09	07	−06	11	33	−05	39	12	08	31
15	Pursuit	56	31	−09	13	01	20	−08	06	17	51
16	Spatial Reasoning	15	16	46	20	14	07	05	38	38	61
17	Figure Classification	01	05	03	06	04	15	16	32	38	30
18	Figure Analogies	10	20	35	18	17	18	40	31	28	60
19	Spatial Visualization I	14	03	34	24	56	15	32	21	18	71
20	Spatial Visualization II	08	17	44	17	42	18	35	36	17	75
21	Decoding	36	12	36	19	01	01	30	37	00	53
22	Number Series Completion	05	47	36	04	16	10	19	00	20	47

[1] Decimal points omitted.
[2] For code numbers see table 7.20.

703120—47—9

that performance in the test is facilitated by quick apprehension of the details of the intersections of lines. It should be noted (see table 28.15) that the test has its highest loading (weighted average) in this factor.

No reasoning test, except Decoding, has a loading greater than 0.15 with this factor. The loading of Decoding implies a need for reduction of this factor by utilizing flag-symbols that are distinctly different from one another.

Rotated factor II is defined by the following tests and loadings:

Test number	Test name	Loading
6	Numerical Operations	0.69
5	Mathematics B	.51
22	Number Series Completion	.47
15	Pursuit	.31

This is the numerical factor, which is also clearly defined in most matrices. Of the experimental reasoning tests, only Number Series Completion appears projected on this factor with a loading greater than 0.20. Comparing the loadings of the Number Series test on this and on the general-reasoning factor (0.36), it can be seen that it is, at least for the aviation-student population, more of a numerical test than a reasoning test.

The substantial loading of the Pursuit test on this factor invites comment. The test does not involve the numerical operations of addition, subtraction, multiplication, and division, but it does involve the location and use of numbers in answering the test items. In two other analyses (see ch. 9 and 16), Pursuit had loadings of 0.19 and 0.26 on the numerical factor. The weighted average loading is 0.25 (see table 28.15). Another test, Organizational Planning (see ch. 9), also has a nonnegligible loading (0.41) on this factor. This test, too, does not involve numerical operations as much as it requires the examinee to locate and remember numbers. Apparently, then, the definition of the numerical factor must be broadened to include more than the simple numerical operations.

Rotated factor III is defined by the following tests and loadings:

Test number	Test name	Loading
5	Mathematics B	0.55
16	Spatial Reasoning	.46
20	Spatial Visualization II	.44
10	Competitive Planning	.41
11	Instrument Comprehension I	.36
21	Decoding	.36
22	Number Series Completion	.36
18	Figure Analogies	.35
19	Spatial Visualization I	.34
4	Reading Comprehension	.33
13	Pattern Comprehension	.33
9	Planning Air Maneuvers	.25
7	Mechanical Principles	.25

This is identified as the general-reasoning factor, usually defined by Mathematics B. All the tests, with the exception of Figure Classification, that were considered to be promising nonverbal reasoning tests contain this factor, but with loadings ranging from only 0.34 to 0.46.

Rotated factor IV is defined by the following tests and loadings:

Test number	Test name	Loading
12	Instrument Comprehension II	0.57
8	Complex Coordination	.52
9	Planning Air Maneuvers	.41
11	Instrument Comprehension I	.39
7	Mechanical Principles	.29
19	Spatial Visualization I	.24

This is the familiar spatial-relations factor, usually best defined by the two leading tests in the tabulation above. No reasoning test has an important loading on this factor; the highest loading is for Spatial Visualization I (0.24).

Rotated factor V is defined by the following tests and loadings:

Test number	Test name	Loading
19	Spatial Visualization I	0.56
7	Mechanical Principles	.54
20	Spatial Visualization II	.42
13	Pattern Comprehension	.50
14	Gottschaldt Figures (QP Part III)	.33

This is the visualization factor. The factor was first defined in this analysis of nonverbal reasoning tests. The tests highest on the factor apparently involve the manipulation of visual imagery. Of the reasoning tests, only the Spatial Visualization tests have loadings on this factor greater than 0.17.

Rotated factor VI is defined by the following tests and loadings:

Test number	Test name	Loading
3	Technical Vocabulary (Navigator)	0.66
4	Reading Comprehension	.65
5	Mathematics B	.34

This is the verbal factor. Its absence in tests utilizing apparently complicated verbal directions is eloquent testimony of the success of careful test-construction. No nonverbal reasoning test has a verbal loading greater than 0.18.

Rotated factor VII is defined by the following tests and loadings:

Test number	Test name	Loading
18	Figure Analogies	0.40
14	Gottschaldt Figures (QP Part III)	.39
12	Instrument Comprehension II	.36
20	Spatial Visualization II	.35
11	Instrument Comprehension I	.34
19	Spatial Visualization I	.32
1	Speed of Identification	.31
21	Decoding	.30
4	Reading Comprehension	.27
13	Pattern Comprehension	.24

This new factor defies precise description. No test has a very high saturation in the factor, and all the tests on it are complex. Two tentative definitions of this factor have been proposed. The first is that it is a visual memory factor, identical with that best defined by the Map Memory tests (see ch. 11), with either or both of these possibly identical with the factor best defined by the Plane Formation tests (see ch. 16). But the presence of Spatial Visualization II and Reading Comprehension, which do not involve pictorial material, is against this interpretation. The second hypothesis is that it is another reasoning factor, although it is not precisely definable yet. Until better evidence is available, this factor may be called reasoning II.

Rotated factor VIII is defined by the following tests and loadings:

Test number	Test name	Loading
16	Spatial Reasoning	0.38
21	Decoding	.37
20	Spatial Visualization II	.36
17	Figure Classification	.32
18	Figure Analogies	.31
19	Spatial Visualization I	.21

All the experimental reasoning tests except Number Series Completion appear projected on this factor, albeit with moderate loadings. There is thus strong indication of the existence of a new reasoning factor. The moderate loadings in this factor, however, make it difficult to formulate a definition. One hypothesis was advanced that all these tests call for the fluent formation of hypotheses. The presence of Spatial Visualization II on the factor, however, is against this interpretation. A reasonable possibility is that all the tests on the factor involve sequential reasoning, i. e., whether "A" is true depends on whether B, C, and D are true. This interpretation emphasizes the evaluation of hypotheses, rather than their formulation. If this is correct, however, the absence of Number Series is difficult to rationalize. A test is urgently needed that will have a high loading on the factor. When that test is constructed, the factor may be defined. Until then, the factor may be called reasoning III.

Rotated factor IX is defined by the following tests and loadings:

Test number	Test name	Loading
16	Spatial Reasoning	0.38
17	Figure Classification	.38
9	Planning Air Maneuvers	.33
10	Competitive Planning	.31
18	Figure Analogies	.28

This factor seems to be integration III (see ch. 10), and perhaps it is clearly described by the phrase "taking into account." In all the tests on this factor, the examinee is required to select one of many courses of action. His selection of the correct one depends upon his ability to take into account all the aspects of the given situation.

Conclusions

This analysis yields no conclusive insight into reasoning tests. The results are a challenge to future investigators.

The main facts and interpretations concerning the experimental reasoning tests may be enumerated as follows:

(1) Reasoning III is probably a true reasoning factor, since all but one of the experimental reasoning tests have moderate loadings on the factor. More research is needed, however, before a clear definition of the ability can be formulated.

(2) That reasoning II is truly another reasoning ability is more dubious. Again, only future research can establish the facts.

(3) The experimental reasoning tests are all factorially complex, and they do not have very high loadings on any factor. Commonly, the tests include variance of the general-reasoning, reasoning II, reasoning III, and integration III factors. All but one of the experimental nonverbal reasoning tests appear to contain the familiar general-reasoning factor, with loadings ranging from 0.34 to 0.46.

(4) Nonverbal-reasoning tests typically are free of even moderate saturations with perceptual, numerical, spatial, visualization, and verbal factors, although several tests have important loadings with one or another of these factors. The loading of Decoding on the perceptual factor points to a defect in test construction, which can be easily rectified in future work with the test. The loading of the Number Series test in the numerical factor is probably unavoidable. There would seem to be little promise of purifying the test to increase reasoning content at the expense of numerical. The visualization content of the Spatial Visualization tests over-shadows their reasoning content. As a matter of fact, in later work with one of these tests (see ch.12) an attempt was made to increase visualization and decrease reasoning content.

(5) Comparing the communalities with the estimates of reliability given in this chapter, it may be seen that for every experimental reasoning test there is considerable undefined nonerror variance. For each test, the approximate percentages of such unknown variance are: 24 percent for Spatial Reasoning; 48 percent for Figure Classification; 22 percent for Figure Analogies; 14 percent for Spatial Visualization I; 12 percent for Spatial Visualization II; 28 percent for Number Series Completion; and 19 percent for Decoding. The weighted average communalities for all analyses (see table 28.15) do not yield significant enlargements of communalities as found in this one analysis, except in the cases of Spatial Reasoning and Figure Analogies.

The analysis sheds light on several other problems. In the first place, the visualization factor was first defined in this analysis. Secondly, the fact that the Pursuit test (in this and other analyses) appears on the numerical factor forces a broadening of the definition of the factor to include noncomputational facility with numbers. Locating, observing, and remembering numbers, as well as adding, subtracting, dividing, and multiplying, are apparently involved. Finally, the appearance of the Pursuit test on the perceptual-speed factor indicates that even apparently relatively unimportant detail discrimination in a test will introduce perceptual variance.

EVALUATION OF REASONING TESTS

While arithmetic-reasoning tests are quite valid against the navigator criterion, they are unsatisfactory in that they are factorially complex. No reasoning test was found that is valid for pilot trainees.

It is noteworthy that no tests have high loadings either on the familiar general-reasoning factor or on the two new factors, reasoning II and reasoning III. It is also noteworthy that, typically, there remain considerable amounts of unknown nonerror variance. Future research should define the new factors and account for the unknown variance. The area of reasoning tests is still largely unexplored.

Conspicuous by their absence are mentions of the concepts of deductive and inductive reasoning. These logical rubrics seem not to yield valid descriptions of psychological factors.

The Complex Coordination Test

The Complex Coordination Test, code number CM701A, a psychomotor test, is mentioned in this chapter and in several later ones, so a very brief description of it is in order here. It is a serial, choice-reaction-time test in which each stimulus is one of 13 spatial patterns of 3 lights each. In systematic correspondence with each stimulus pattern, the correct response is a unique adjustment of imitation stick-and-rudder controls. Each correct reaction automatically brings a new stimulus. The score is the number of reactions completed in 8 minutes.

CHAPTER EIGHT

Judgment Tests[1]

INTRODUCTION

Judgment in Aviation

Judgment has been one of the major areas of research in aviation psychology. An important reason for this was the common practice of flight instructors to place errors of perception, visualization, and reasoning all under the broad category of "poor judgment."

Poor judgment is one of the most frequently mentioned reasons given by instructors for eliminating cadets from flying training. It was mentioned in 50 percent of the cases in an analysis of the reasons stated in faculty-board proceedings for eliminating 1,000 aviation cadets from elementary pilot training in the latter part of 1941. Some typical comments classified under the category of poor judgment are: "dangerous judgment in traffic"; "unable to make sound decisions in traffic or in the vicinity of other planes"; "choice of fields and judgment in simulated forced landings has been weak"; "unable to exercise safe judgment in the air"; and "fails to discriminate safe from unsafe flying."

In an attempt to clarify the concept of judgment, aviation psychologists asked flight instructors to define what they meant by judgment. Some typical definitions constructed from the comments of flight instructors are:

1. The ability to react immediately and appropriately to stimuli with which an individual is unacquainted.
2. "Headwork" or the ability to react correctly without deliberation, or the ability to fly without confusion in traffic and under unusual circumstances.
3. Knowledge, plus speed of reaction, plus freedom from emotional confusion.
4. The ability to react appropriately in a surprise situation.
5. Ability to grasp the situation as a whole, not being absorbed with minor details.

Previous Studies of Pilot Judgment

During World War I tests of judgment of distance, speed, and time were used. These included estimation of length of sticks, of the relative speeds of four revolving disks, of the time required for sand to flow from one container to another, and of the curves and relative speed of two white spots moving along converging lines in a horizontal plane (1).

[1] Written by Staff/Sgt. Benjamin Fruchter.

In 1940, Kelly had flight instructors rate 110 civilian pilot training students on a 14-item graphical scale of pilot competency (3). Thirteen of the items were intercorrelated and analyzed by Thurstone's centroid method. Three factors were found necessary to account for the intercorrelations and were identified as (1) skill, (2) judgment, and (3) emotional control. The items that had significant loadings on the judgment factor are as follows:

How good is his judgment with regard to taking flying risks (weather, stunting, etc.)?

Does he show respect for a ship and its motor?

How well is he satisfied with his flying ability?

Is he inclined to show off while flying a plane?

How carefully does he check his plane and engine before taking off?

Judgment in the AAF Qualifying Examination, AC10A

The first form of the AAF Qualifying Examination, AC10A (see Report No. 6), introduced in January 1942, included a number of judgment-type items like the following:

A pilot has to make a forced landing near a mountain cabin. He finds that the nearest phone is at an isolated fire ranger's cabin 14 miles across the mountains to the north. It is winter. He sets out on foot for the ranger's cabin at 6 a.m., carrying food for only one meal. At 10 a.m., having met no one, he comes to three branches of the trail, all unmarked. His most practical decision would be to:

A. Follow the trail which appears to lead in the right direction until he reaches the cabin or the end of the trail.

B. Turn back immediately to his starting point.

C. Leave the trail and go due north by compass.

D. Walk along the trail which appears to lead in the right direction until noon, then turn back if not sure of his location.

E. Stay in the fork in the trail and wait for someone to come by.

The judgment subtest (15 items) of AC10A had a reliability of 0.36, based on 370 pilot students. It had a biserial correlation of 0.36 with the criterion of graduation-elimination from primary pilot training, based on 545 cases in class 42G. These facts seemed to substantiate the hypothesis that practical judgment was a measurable psychological category and to make desirable further analysis of the problem.

Research on Judgment Tests for Classification

Research on judgment attempted (*a*) to analyze judgment tests and (*b*) to analyze the concepts of practical judgment as described by instructors. Both lines of attack were fruitful, but at the war's end neither had been exhausted, by any means.

In the rest of this chapter the complex nature of the tests is demonstrated, and their unique contribution—a judgment factor—is revealed. The role of background information in both judgment tests and the pilot criterion was fairly well verified, and certain types of information (for example, mechanical information) were found to be contributory to success both in test performance and in primary pilot training.

The relations of judgment to reasoning and perceptual abilities are pointed out in this and in other chapters. The hypothesis of a functional thought-fluency factor is mentioned, but conclusions cannot be reached owing to the lack of final results.

PRACTICAL JUDGMENT TESTS

Practical Judgment, CI301BX1 [3]

The hypothesis was formed that a major part of the variance of typical judgment-test items is attributable to individual differences in pertinent informational background. This grew out of a conviction that the judgment items in AC10A have a very high mechanical-knowledge component and that this accounts in large part for the validity of the test. The general hypothesis can be tested by a study of the mechanical-information type of item alone. Such items were expected to correlate with Mechanical Principles, CI903B, and Mechanical Information, CI905B, tests whose mechanical content is known.

Description. (1) *Internal characteristics.*—Test CI301BX1 contains 80 items and covers a large range of situations which are considered solvable by ordinary judgment. Twenty-eight items are considered to be dependent primarily upon a knowledge of mechanics, and 52 items are considered to be essentially non-mechanical. The items are segregated into two parts, each composed of random halves of the mechanical and nonmechanical items. The following are examples of a nonmechanical and a mechanical judgment item respectively:

An officer must send an important confidential message about 4 miles through enemy lines into an area which is very closely guarded. It is important that the message not fall into enemy hands, but it is equally important that the message get through. Under the circumstances it would be best for him to:

A. Write out the message and give it to a runner with orders to get through as quickly as possible.
B. Send one runner with a decoy message with instructions to get through as soon as possible.
C. Send one runner with the written message and another to act as his guard.
D. Write out duplicate forms of the message and give them to two runners with instructions to get through as soon as possible.
E. Have two runners memorize the message and instruct them to get through as soon as possible.

You are operating a large water-cooled motor with a heavy load, when you notice that a bearing is heating excessively. It would be best for you to:

A. Stop the motor immediately and lubricate the bearing.
B. Remove the load, lubricate the bearing freely, with the motor running slowly.
C. Run the motor slowly with the load.
D. Continue to operate the motor at present speed and lubricate the bearing.
E. Stop the motor immediately and add cold water.

[3] Developed in the Office of the Air Surgeon, Headquarters, Army Air Forces, and Psychological Research Unit No. 1.

(2) *Administration.*—Directions are printed in the booklet, and the test is largely self-administering. The time allowed for part I is 40 minutes and for part II, 40 minutes. It has been standard practice to allow 1 minute per item for this type of test.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—Some statistics were computed separately for the 28 mechanical items and the 52 nonmechanical items, and some for all items together. All the data reported are for examinees tested in December 1942 at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 8.1.

TABLE 8.1.—*Distribution constants for Practical Judgment, CI301BX1, based upon a sample of 202 unclassified aviation students*

Score	M	SD
Mechanical judgment	14.8	4.4
Non-mechanical judgment	25.6	5.7

(2) *Internal consistency.*—Analysis of responses of sample groups yielded the internal-consistency data given in table 8.2.

TABLE 8.2.—*Internal-consistency data for Practical Judgment, CI301BX1, based upon highest and lowest 27 percentages of a sample of 202 unclassified aviation students*

Criterion	Items	Mϕ	SDϕ	Range of ϕ low	Range of ϕ high
Mechanical score	Mechanical	0.31	0.11	0.02	0.49
Non-mechanical score	Non-mechanical	.25	.11	.04	.46

(3) *Reliability coefficient.*—By the alternate-forms method (part I vs. part II), an estimated reliability coefficient of 0.62, corrected for length, was obtained. This figure is based on a sample of 202 unclassified aviation students.

(4) *Difficulty.*—Based upon item analysis of the responses of 500 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.53, corrected for chance, with a range of 0.00 to 0.91 and a standard deviation of 0.24.

(5) *Factorial composition.*—The most significant loadings for the mechanical-items score are in the mechanical-experience (0.54), judgment (0.36), and visualization (0.29) factors; and for the nonmechanical score, in the judgment (0.39), planning (0.36), and visualization (0.30) factors. The mechanical-experience loading in the nonmechanical items is only 0.13, which is in line with the hypothesis that some items contain mechanical-information content and some do not; and, that by design and selection the two types can be fairly well segregated. The communalities for the two types of items are 0.59 and 0.48 respectively. For a fuller picture of the factorial composition of this test see appendix B.

(6) *Test validity.*—Validation results are given in table 8.3.

TABLE 8.3.—*Validity data for Practical Judgment, CI301BX1, based upon graduation--elimination from primary training of a sample of 267 pilots*

Score	M_g	M_e	SD_t	r_{bis}
Mechanical judgment	15.15	14.90	3.57	0.04
Non-mechanical judgment	25.76	25.30	5.44	.05

(7) *Item validity.*—Validation of items of both mechanical judgment and nonmechanical judgment items combined revealed a mean phi of 0.00, based upon the responses of 200 graduates and 64 eliminees from training. The standard deviation of phi values is 0.20 and the range is from −0.30 to 0.18.

Evaluation.—An examination of the factor loadings of the two types of judgment items supports the hypothesis that a large part of the variance of the mechanical-judgment items is accounted for by individual differences in mechanical information. The two types of items have approximately equal weights on the visualization factor and also on a factor identified as judgment. In either case, however, the variance in judgment is only approximately 15 percent. Assuming that this factor is weighted 0.10 for the pilot criterion, for which there is some evidence, we should expect a test of this type to add somewhat to the validity of a pilot battery of which it is a part.

From the factorial composition of the two tests we should expect (see chapter 28) a validity of 0.28 for the mechanical items (to be compared with the 0.36 for the judgment test in AC10A) and 0.17 for the nonmechanical items. The pilot-validity figures given in table 8.3 are by no means typical of obtained validities for these kinds of items. Weighted averages of a number of estimates of validities derived from similar forms are 0.18 and 0.13, respectively. From these results we can be fairly satisfied that all factors with pilot validity are known in these types of judgment tests. From a comparison of reliabilities and communalities we can conclude that all common factors, valid and invalid, are accounted for.

Practical Judgment, CI301BX2

This test is a 40-item revision of CI301BX1. The items of the previous form having mechanical content are eliminated. This was done on the basis of a factor analysis of the previous form in a matrix containing selected tests from the classification battery. This analysis indicates that nonmechanical judgment items, in addition to mechanical and intellectual loadings, define a new factor tentatively characterized as a judgment factor. The revision was made in an attempt to reduce the mechanical loadings and increase the loading on the new factor.

Description. (1) *Internal characteristics.*—A sample item follows:

The principal reason why barracks in Army camps are built according to an identical plan is that this method:

A. Requires the least use of construction materials.
B. Makes possible the greatest speed of construction.
C. Allows construction of barracks that will last a long time.
D. Requires a low cost of up-keep of the barracks after completion.
E. Results in a military appearance which is similar in all Army camps.

(2) *Administration.*—The two parts are timed separately, with an allowance of 20 minutes per part.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data reported below are for unclassified aviation students tested in February 1943 at Psychological Research Unit No. 3. Those who entered primary pilot training were in class 43K.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 8.4.

TABLE 8.4.—*Distribution constants for Practical Judgment, CI301BX2*

Group	N	M	SD
Unclassified aviation students	571	11.7	4.2
Classified pilots	438	10.9	4.4

(2) *Internal consistency.*—The degree of homogeneity of the items of the test is indicated by a mean internal-consistency phi of 0.25, a standard deviation of the phi distribution of 0.10, and a range of values from −0.11 to 0.48. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 360 unclassified aviation students.

(3) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.43, corrected for length, was obtained. This figure is based on a sample of 485 classified pilots.

(4) *Difficulty.*—Based upon item analysis of the responses of 480 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.33, corrected for chance, with a range from 0.00 to 0.88 and a standard deviation of 0.20.

(5) *Test validity.*—A sample of 571 pilots yielded a biserial correlation of −0.02, corrected for restriction of range, between performance in this test and the graduation-elimination criterion in primary training. The mean score for graduates was 11.70, for eliminees 11.86, and the standard deviation for both combined was 4.20. Of this sample 85 percent were graduates, and the standard deviation assumed for the unrestricted pilot stanine distribution was 2.00.

(6) *Item validity.*—Validation of items revealed a mean phi of 0.00, based upon the responses of 500 graduates and 100 eliminees from primary training. The standard deviation of phi values is 0.07, and the range is from −0.11 to 0.15.

Reasoning content of Practical Judgment, CI301BX2.—The hypothesis was adopted that the examinees achieved correct solutions to many of the items in this form by reasoning. To test this hypothesis, it was decided to do an item analysis using the same 360 cases that were used in the internal-consistency item analysis, basing the phi coefficients this time on the criterion of Arithmetic Reasoning scores, which best defines the general reasoning factor. The product-moment correlation coefficient between the internal-consistency phis and those based upon the arithmetic-reasoning criterion is 0.56. This is considered corroboration of the hypothesis that Practical Judgment Test, CI301BX2, contains considerable reasoning content.

Evaluation.—This attempt to enhance measurement of the judgment factor found in the factorial analysis of the Foresight and Planning I battery (see ch. 9) was considered unsuccessful, inasmuch as the items are largely solvable by a type of reasoning already measured by the Arithmetic Reasoning Test, CI206C. This factor has zero validity for pilot selection.

Practical Judgment, CI301BX3 [a]

This test is a revision of Practical Judgment, CI301BX2. It represents another attempt to clarify the nature of the judgment factor.

Description. (1) *Internal characteristics.*—Part I contains the 25 items from the previous form having internal-consistency phis of 0.25 or higher. Part II contains 25 newly-constructed items of the work-planning type which seemed most likely to measure the new factor.

The following is a sample of the type of item in part II:

At a mobile army encampment, it is necessary to use buckets to fight a fire 100 feet away from a stream and up a hill. Fifteen 5-gallon buckets and 60 men are available. The best procedure would be to:

A. Select 15 men and have each run between the stream and the fire carrying buckets, and replace the entire crew at intervals.

B. Line up the 60 men from the building to the stream and pass the buckets from man to man from the stream to the building, the last man throwing the water on the fire and returning with the empty bucket.

C. Make one line of 25 men to pass the buckets from the stream to the building with one line of 10 men to throw the empty buckets back: 2 men to dip, 2 men to throw water on the fire, and the rest for relief.

D. Detail 2 men to dip and 2 to throw water on the fire; assign 26 to carry buckets from the stream to the building and back; replace the entire crew with 30 fresh men in 15 minutes.

E. Detail 5 men to fill the buckets at the stream and leave them at the bank, where they can be picked up by the rest of the men who will carry them to the building and return with empty buckets.

(2) *Administration.*—The time is 25 minutes each for Parts I and II.

(3) *Scoring.*—The scoring formula for this test is R−W/4.

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter, Jr. and Lt. Frank J. Dudek.

Statistical results.—All results reported below are for examinees tested at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—A sample of 242 unclassified aviation students in May 1943 yielded a mean score of 17.2, a standard deviation of 5.3, and a range from 3 to 34.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.28, a standard deviation of the phi distribution of 0.12, and a range of values from −0.05 to 0.46. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 480 unclassified aviation students tested in July 1944.

(3) *Reliability coefficient.*—Estimates of reliability computed for the two part-scores for a sample of pilots are given in table 8.5.

TABLE 8.5.—*Estimated reliability coefficients by the odd-even method for Practical Judgment, CI301BX3, based upon a sample of 167 classified pilots*

Score	r'_{II}	r_{II}
Part I	0.37	0.54
Part II	.26	.41

(4) *Difficulty.*—In an item analysis of the responses of 480 unclassified aviation students tested in July 1944, the test yielded a mean proportion of correct responses of 0.32, corrected for chance, with a range from 0.00 to 0.78, and a standard deviation of 0.22.

(5) *Factorial composition.*—Part I and part II of the test were treated separately in analyzing this test. The most significant loadings for part I are in the judgment (0.30), verbal (0.30), visualization (0.29), and planning (0.28) factors. The most significant loadings for part II are in the judgment (0.45), general reasoning (0.40), and mechanical-experience (0.32) factors. Part I has a loading of 0.03 in the general-reasoning factor, and part II has a loading of 0.03 in the verbal factor and only 0.08 in the planning factor. The communalities are 0.49 and 0.51 for parts I and II, respectively. For a fuller picture of the factorial composition of this test see appendix B.

(6) *Item validity.*—Validation of items revealed a mean phi of 0.01, based upon the responses of 200 graduates and 64 eliminees from primary pilot training (class 43J). The standard deviation of phi values is 0.07, and the range is from −0.28 to 0.19.

Evaluation.—The newly-constructed items in part II of this test have a loading of 0.45 on the judgment factor, compared with a loading of 0.30 for the items in part I, which were taken from previous forms of the test. It may therefore be concluded that items of the work-planning type best define the judgment factor. Reliabilities of scores continue to be low, though communality of items cannot be doubted. The existence of a judgment factor seems to be fairly well verified, but its exact nature needs further clarification. The type of test in which it can be very strong and unique is yet to be found. It is present, however, in tests that are not designated as judgment tests, so it is possible that some new type of test can be designed to improve its measurement.

Practical Judgment, CI301C [•]

This form of the test was compiled from previous forms for inclusion in the September 1944 classification battery. The items were selected on the following bases: 1. validity; 2. internal consistency; 3. full coverage of field; 4. nonmechanical content; 5. difficulty level.

A sample item is:

A man on a very urgent mission during a battle finds he must cross a stream about 40 feet wide. A blizzard has been blowing and the stream has frozen over. However, because of the snow, he does not know how thick the ice is. He sees two planks about 10 feet long near the point where he wish to cross. He also knows where there is a bridge about 2 miles downstream. Under the circumstances he should:

A. Walk to the bridge and cross it.
B. Run rapidly across on the ice.
C. Break a hole in the ice near the edge of the stream to see how deep the stream is.
D. Cross with the aid of the planks, pushing one ahead of the other and walking on them.
E. Creep slowly across on the ice.

(1) *Administration.*—Thirty minutes are allowed for the 30 items.

(2) *Scoring.*—The scoring formula is $2R - 2W/3$.

Statistical results.—Only distribution statistics were available at the time this was written. For 2,917 unclassified aviation students tested at Medical and Psychological Examining Units, the mean score was 20.0, and the standard deviation 8.6.

Practical Judgment, CI301DX1 [•]

The items in this form were constructed to provide a reserve pool of validated judgment items for future revisions of test CI301C.

PRACTICAL ESTIMATION TESTS

These tests were constructed primarily for the purpose of analyzing the informational background of judgment tests. It was noted that numerous items in the Practical Judgment tests require the examinee to make estimates of the weights of certain objects, of the amounts of time necessary to carry out certain tasks, etc. It was deemed desirable to discover how important this content was in defining the judgment factor. If this ability should prove to be the unique component of judgment tests, it would be possible to construct purer tests of judgment.

[•] Developed at Psychological Research Unit No. 3.
[•] Developed at Psychological Research Unit No. 2. Chief contributors: Cpl. Robert E. Lambert, Jeanne L. Lipman, Sgt. Robert B. Porter.

Practical Estimations, CI308AX1 *

Description. (1) *Internal characteristics.*—The 86 items of this test are divided into 4 parts. Parts I and II call for relative judgments; parts III and IV, for absolute judgments. Part I contains items in each of which the amounts of time required in five different situations are compared. In part II five distances or sizes of five different objects are compared in each item. In part III questions are asked about the amount of time required for a particular activity, and the answer is selected from a scale of 15 steps. In part IV questions about the length and sizes of objects are asked, and the answers are selected from a scale of 15 steps. Two representative problems are:

Which of the following could be done in the shortest time?

- A. Walking 5 miles on snowshoes.
- B. Riding a horse 18 miles.
- C. Swimming 2 miles.
- D. Rowing a boat across a lake 2½ miles wide.
- E. Walking 6 miles on flat terrain.

Of the following, which is the shortest?

- A. Six building bricks laid end to end.
- B. The length of wire used in making a wire coat hanger.
- C. The distance from the floor to the door knob.
- D. The width of the average door.
- E. Four sheets of typing paper laid end to end.

(2) *Administration.*—The time limit for part I is 12 minutes; for part II, 9 minutes; for part III, 15 minutes; and, for part IV, 11 minutes.

(3) *Scoring.*—The scoring formula for parts I and II is R−W/4; for parts III and IV, R−W/5.

Statistical results.—The data reported below are for examinees tested in April and May 1943 at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—A sample of 237 unclassified aviation students yielded a mean total score of 29.8, a standard deviation of 7.1, and a range from 14 to 47.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.13, a standard deviation of the phi distribution of 0.11, and a range of values from −0.01 to 0.39. These statistics are based upon analysis of the responses of the highest 27

TABLE 8.6.—*Reliability coefficients for part scores of Practical Estimations Test, CI308A, computed by the odd-even method, based upon a sample of 183 unclassified aviation students*

Score	r'_{11}	r_{11}
Part I	0.27	0.42
Part II	−.03	−.05
Part III	.07	.13
Part IV	.26	.42

* Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Frank J. Dudek and Lt. Linn Hutchinson.

percent and the lowest 27 percent in total score of a group of 750 unclassified aviation students.

(3) *Reliability coefficient.*—Estimates of reliability computed for each part of the test are given in table 8.6.

(4) *Difficulty.*—Based upon item analysis of the responses of 750 unclassified aviation students, the test (all four parts combined) yielded a mean proportion of correct responses of 0.34, corrected for chance, with a range from 0.02 to 0.71 and a standard deviation of 0.16.

(5) *Factorial composition.*—The scores for part I and part II were treated separately for factor-analysis purposes. The most significant loadings for part I are in the planning (0.36), judgment (0.36), and mechanical-experience (0.33) factors. The most significant loadings for part II are in the mechanical-experience (0.32), planning (0.31), and numerical (0.28) factors. The numerical loading in part I is 0.13, and the judgment loading in part II is 0.02. The communalities are 0.39 and 0.35, respectively. The first one is very close to the estimate of reliability, and the second one clearly shows that the reliability of part II was grossly underestimated. For a fuller picture of the factorial composition of this test see appendix B.

Items calling for absolute judgments had low communality with estimation items calling for relative judgments, with Practical Judgment, CI301BX1, and with selected experimental and classification tests. They were, therefore, not included in the matrix for factor analysis.

Evaluation.—Items calling for relative judgments, though having low reliability, gave indication of having substantial communality with the experimental battery with which they were administered. It is interesting to note that of the two parts of the Practical Estimations Test, CI308AX1, that were analyzed with the Foresight and Planning II battery (see ch. 9), the relatively complicated judgments of part I (involving estimations of distance and time) have a significant loading (0.36) on the judgment factor, whereas the relatively simple judgments of part II (involving judgments of distance only) have an insignificant loading (0.02).

The pilot validities to be expected from the factor compositions are 0.15 and 0.14 for parts I and II, respectively (see table 28.18). The average validities found for the comparable form, CI308BX1 (see ensuing discussion), are 0.14 and 0.13, which check very closely. These data and the comparison of communalities and reliabilities indicate that all common factors, valid or invalid, are accounted for.

Practical Estimations, CI308BX1 [1]

This test is a revision of Practical Estimations, CI308AX1. The items in the previous form calling for absolute, rather than relative, judgments had so little communality that they were dropped from this form. This

[1] Developed at Psychological Research Unit No. 3. Chief contributors: Cpl. Robert E. Lambert, Pvt. James A. Walker.

783130—47—10

form incorporates the items from the previous form calling for relative judgments, plus newly-constructed items of the same nature.

Description. (1) *Internal characteristics.*—Part I contains judgments involving time and distances. Part II contains judgments involving distances.

The following items illustrate the types in parts I and II respectively:

Which of the following travels fastest?

A. A batted baseball as it leaves the bat.
B. A polo ball as it leaves the mallet.
C. An arrow as it leaves the bow.
D. A tennis ball just after the serve.
E. A prizefighter's fist in the middle of the swing.

Which of the following is most nearly the same as the distance from one side of the car windshield to the other side?

A. The distance of a car doorhandle above the ground.
B. The distance of a doorknob above the house floor.
C. The length of an unfolded newspaper.
D. The width of an average door.
E. The width of an ordinary desk.

(2) *Administration.*—The two parts are timed separately, 23 minutes being allowed for the 35 items in each part.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data reported below are for examinees tested at Psychological Research Unit No. 3 in June and July 1944.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 8.7.

TABLE 8.7.—*Distribution constants for Practical Estimations, CI308BX1, based upon a sample of 750 unclassified aviation students*

Score	M	SD
Part I	12.2	3.8
Part II	10.8	4.3

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.13, a standard deviation of the phi distribution of 0.10, and a range of values from −0.05 to 0.39. These statistics are based upon the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 750 unclassified aviation students.

(3) *Difficulty.*—Based upon item analysis of the responses of 400 unclassified avia[illegible]dents, the test yielded a mean proportion of correct responses of [illegible] [illegible]rrected for chance, with a range from 0.00 to 0.75 and a standard deviation of 0.16.

(4) *Test validity.*—Validation data are shown in table 8.8.

TABLE 8.8.—*Validity data for Practical Estimations, CI308BX1, based upon samples of pilots, using the graduation-elimination criterion*

Part	Score	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
I	Rights	[2]660	0.80	16.52	15.98	3.29	0.10	0.14
II	"	[2]660	.80	15.34	15.34	3.47	.00	.07
I	Wrongs	[2]660	.80	17.12	17.61	3.71	−.08	−.11
II	"	[2]660	.80	18.41	18.05	3.87	.06	−.01
I+II	Rights	[2]660	.80	31.86	31.32	5.12	.06	.14
I+II	Wrongs	[2]660	.80	35.54	35.66	6.01	−.01	−.07
I	Rights	[3]623	.82	16.61	15.93	3.45	.11	.15
II	"	[3]623	.82	15.75	14.80	3.60	.15	.19
I	Wrongs	[3]623	.82	17.53	17.99	3.76	−.07	−.12
II	"	[3]623	.82	18.24	19.21	3.95	−.14	−.18
I	R—W/4	[3]623	.82	12.23	11.43	4.18	.11	.15
II	R—W/4	[3]623	.82	11.19	10.00	4.40	.15	.19
I+II	Rights	[3]623	.82	32.36	30.73	5.27	.18	.22
I+II	Wrongs	[3]623	.82	35.77	37.20	6.11	−.13	−.19

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Tested in the period from June 6 to July 6, 1944.
[3] Tested in the period from July 3 to October 7, 1944. Does not overlap sample of 660.

Evaluation.—The moderate pilot validity of this test indicates that the judgment factor has low to moderate validity for the prediction of success in pilot training, after validities due to mechanical-experience and planning factors are taken into account.

Evaluation of Practical Estimation Tests

An attempt was made to discover whether tests calling for estimations of time and distance would best define the judgment factor. It was concluded that items that call for complicated, relative estimations have a significant loading in the judgment factor. Practical estimation tests are not, however, a pure measure of the judgment factor. In addition to their loadings with the judgment factor, they also have significant loadings with the planning and mechanical experience factors. Moreover, they were found to be less heavily weighted with the judgment factor than are practical-judgment items of the work-planning type.

FLUENCY TESTS

Another aspect of the concept of practical judgment that was considered worth investigating is the ability to call to mind experiences and hypotheses that are of aid in solving a practical problem. It is supposed that the ease of evocation of pertinent facts and hypotheses is an important element in solving problems which do not lend themselves to conventional solutions.

Thurstone (4) in his factor analysis of 57 tests identified two verbal factors. The first was defined by tests in which the subject deals with ideas and meanings of words. He called this factor "verbal relations." The second was defined by tests in which the subject recalls single and isolated words. He called this factor "fluency in dealing with words."

It was hypothesized that there is a more general fluency factor, not confined to the recalling of words, which facilitates thinking toward the solution of problems. It was intended to construct a number of tests of fluency and to submit the hypothesis of one general fluency factor to the test of factor analysis.

The measurement of fluency requires the use of unusual problem situations to which the examinee is required to respond with the enumeration of as many responses as occur to him in a given amount of time. On the one hand, the multiple-choice type of item would seem to defeat one's purpose, since providing the examinee with ready responses makes recall on his part unnecessary. At the same time, the standard multiple-choice item seems to be indicated in order to avoid such irrelevant factors as verbosity and speed of writing. Machine scoring is also almost a compulsion in air-crew testing. The need for satisfying both demands—freedom of response and machine scoring—presented many difficulties.

As a group, fluency tests were developed late in the war. Some were completed in time for administration for intercorrelation studies, but none in time for validation.

Verbal Recognition, CI322A [a]

This test is based upon the assumption that an examinee's ease in evoking solutions to practical problems is related to the ease with which he can unscramble words that belong to a named category.

Description. (1) *Internal characteristics.*—The approach employed is to name a category and then present 10 "scrambled words." The categories are (1) animals, (2) building materials, (3) sports, (4) men's clothing, and (5) means of transportation. The first letter of each scrambled word is capitalized, regardless of where it appears in the scrambled word. It is believed that by capitalizing the first letter of each word the examinee is forced to think of various alternatives in the given category starting with the capitalized letter rather than the more trial-and-error task of unscrambling each word. Perceptually difficult items were chosen with the expectation that the examinee would prefer to resort to verbal fluency to solve them. The words in part I are scrambled haphazardly. The same five categories appear in part II, but the scrambled words are presented with all the consonants arranged alphabetically, followed by the vowels presented in the same way.

(2) *Administration.*—The following are the directions for the test:

This is a test of your ability to recognize the names of things quickly and accurately when the letters in the names have been mixed up. Two things will assist you in figuring out what the scrambled words are:

1. The first letter of each word will be capitalized, regardless of where it appears in the scrambled word.
2. The 10 words on a page are all names of the same kind of things, such as colors in the sample problems below.

Colors

1. deR.
2. nageOr.

The answer to problem 1 is Red. The answer to problem 2 is Orange.

When the test begins, write the words you are able to unscramble easily on the separate work sheet provided. Be sure to write each word after its corresponding

[a] Developed at Psychological Research Unit No. 2. Chief contributor: Sgt. David Grossman.

number and make the words readable. Misspelling of words will not be counted against you, however.

This is a speed test. If you cannot unscramble a word quickly, go on to the next. You must work rapidly, since you will be allowed only 1¼ minutes to unscramble the 10 words in each group. Do not go on to the next group until the signal is given.

When the test is completed, the test booklet is collected, but the work sheet is retained by the examinee. A specially prepared IBM answer sheet is given him together with the following oral instructions:

The names of the five groups that appeared in the test booklet appear at the top of this answer sheet. On your work sheet the answer to item 1 should be Lion. Now look under the first column on the answer sheet headed Animals and find Lion. It is the fifth word from the top. Blacken in the first space to the left under Lion. Do this for each of the answers that you have on your work sheet.

The time limits are 6¼ minutes for part I and 6¼ minutes for part II.

(3) *Scoring.*—The score is the number of correct responses.

Verbal Recognition, CI322B [9]

Variations of the test.—It is similar to form A except that the examinee must consider all five categories simultaneously rather than just one at a time, as in the previous form. At the top of each page of the test booklet the five categories appear opposite the letters A through E. The examinee indicates the category to which the unscrambled word belongs by blackening the appropriate space on the IBM answer sheet. This obviates the need for using a work-sheet and later transcribing the answers to a specially prepared IBM answer sheet for machine scoring. This form of the test is believed to be more difficult, because there are five reference categories for each response.

Similarities Test, CI319A [10]

This is a test of the ability to recall quickly previously acquired information about common objects.

Description. (1) *Internal characteristics.*—Pairs of common objects are listed in a workbook, each followed by 15 spaces in which the examinee is directed to list ways in which 2 objects are alike. An attempt was made to minimize the verbal factor by (1) limiting the number of words per response and (2) employing relatively simple material.

(2) *Administration.*—The following are the directions:

This is a test to see how quickly you can think of ways in which different objects are alike. In this booklet 20 pairs of objects are presented, 10 pairs in part I and 10 pairs in part II. Under each pair are lettered spaces in which you will write down as many ways as you can think of in which the objects are alike.

Now look at the sample below. Several similarities have been listed in the spaces provided to show you how to enter your answers. A sample problem follows:

---

[9] Developed at Psychological Research Unit No. 2. Chief contributors: Sgt. David Grossman, Capt. John I. Lacey.

[10] Developed at Psychological Research Unit No. 2. Chief contributor: Tech./Sgt. Paul C. Davis.

Sample problem:

Apple and Orange are alike:

A. sweet		
B. round		
C. colored		
D. have	seeds	
E. fruit		
F. have	skins	
G. grow	on	trees
H.		
I.		
J.		
K.		
L.		
M.		
N.		
O.		

Notice that the similarities listed concern real characteristics of the objects, such as structure, use, or operation. Such statements as "bought in stores" and "cost money," which do not describe the objects are not acceptable as answers. Also note that "both" is assumed and need not be written down. As indicated by the above sample, you may use not more than three words in describing any similarity.

The items are timed separately, 1 minute per item. There are 10 items in part I and 10 items in part II. When both parts of the test have been completed, the examinee is given the following oral directions for recording his answers on a 15-place IBM answer sheet:

We will now record the answers on the answer sheet. Look at the front of the test booklet. The sample item, number 26, has answers listed in all the spaces up to and including H. Find number 26 at the top of the right-hand column on the answer sheet and draw a solid line through all the spaces from A through H opposite it. (Illustrate on board.) Do this now. Now begin with item number 1 and record your answers in this manner in the proper spaces on your answer sheet. Work as rapidly as you can.

(3) *Scoring.*—The score is simply the number of responses, one unit of credit being allowed for each similarity written down. This assumes that quality of responses is irrelevant. It is intended to test this assumption before using this numerical index in factor analysis.

Word Association Test, CI318A [11]

This test measures two assumed aspects of fluency: (1) rapidity of association and (2) ease of change of set. Since there were no prospects for its administration, the test was not printed. Its description, however, may be of interest.

Description. (1) *Administration.*—The following are from the directions for the test:

[11] Developed at Psychological Research Unit No. 2. Chief contributors: Staff/Sgt. Arthur Z. Cerf, Lt. Cecil H. Patterson.

This is a test of your ability to recognize associations between words. A key word will be given which may have several meanings and may be associated with one or more of five words opposite it. Your task is to select the words that seem to have similar meaning or are closely associated with the key word, and to blacken the corresponding spaces after the item numbers.

Look at Sample Problem 1:

Sample problem 1

Key word	A	B	C	D	E
Order	neat	tried	command	purchase	single

Alternates A, C, and D are correct. A—neat, means orderly; C—command, to give an order; D—purchase, to order material. Blacken spaces A, C, and D after number 1 on your answer sheet.

In taking the test, here are the things to remember:

1. There may be 1, 2, 3, 4, or 5 correct responses to each item.
2. Correct answers are those which have the same meaning or are closely associated with the key word. Some slang expressions, which may be correct, will be included.
3. Words which sound the same as the key word but have a different meaning are considered wrong responses.

The time limits for the test are: part I, 5½ minutes, part II, 5½ minutes.

(2) *Scoring.*—The scoring formula for the test is R—W/2.

Camouflaged Words, CI323A[12]

This test is a modification of Thurstone's Mutilated Words. Two factors are thought to be measured by this test: (1) ease of evocation of hypotheses (fluency) and (2) changeability of set.

Description.—The items were pretested, and an attempt was made to secure items of approximately 0.5 difficulty with 2 seconds permitted per response. Each mutilated word has two items based upon it.

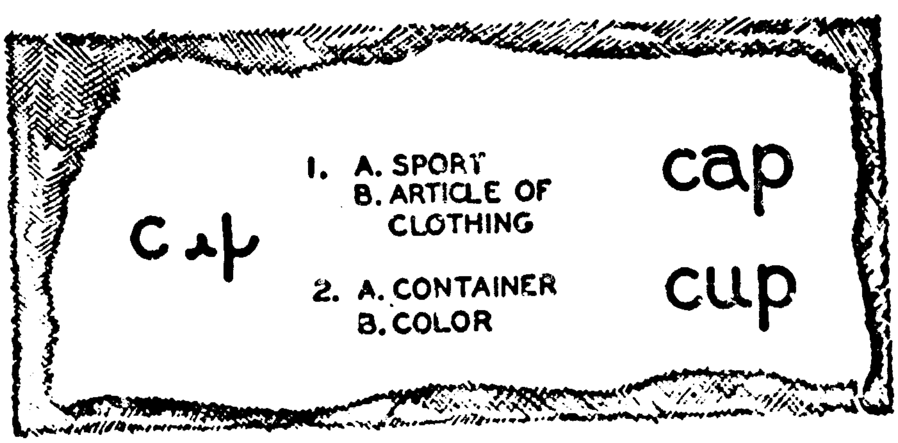


FIGURE 8.1
SAMPLE ITEM OF CAMOUFLAGED WORDS, C1323A

[12] Developed at Psychological Research Unit No. 2. Chief contributors: Florence R. Grossman, Capt. John I. Lacey.

(1) *Internal characteristics.*—The test is divided into 2 comparable parts of 30 problems each (based on 15 mutilated words in each part)

(2) *Administration.*—The following are from the directions to the test:

Pilots must be able to detect and identify camouflaged objects. The purpose of this test is to determine how well you can identify words, parts of which have been removed or camouflaged.

Look at the sample problems below. (See figure 8.1.)

Both sample problems above are based on the camouflaged word at the left. Two categories, A. Sport, and B. Article of Clothing, are listed for problem 1. The word that has been camouflaged is the name of a Sport or the name of an Article of Clothing. Your task is to:

a. Think of words you know in each category until you discover the word that has been camouflaged.

b. Mark your answer sheet A or B according to the correct category.

The same procedure is repeated for problem 2 with two new categories.

The time limits for the test are as follows: practice problems, 3 minutes; administration, 5 minutes; part I, 10 minutes; part II, 10 minutes; total time, 28 minutes.

(3) *Scoring.*—The scoring formula for this test is R—W.

Ambiguous Ink Blots, CI317A [13]

This test is designed to measure the speed with which varied responses can be evoked from constant stimuli consisting of ink blots. The test was not printed, since there were no prospects for its administration.

Description. (1) *Administration.*—The following are from the directions to the test:

This is a test to determine the number of objects you can find easily in a mass of blots and lines which at first may appear to be meaningless. You will be shown a picture, followed by the names of 15 objects. Your task is to study the picture and indicate which ones of the 15 objects you can see in the picture.

Look at sample problem 1, and decide which of the objects A through O can be found in this picture. For each object that you can find, blacken the appropriate space on your answer sheet. Do this now. (See Figure 8.2)

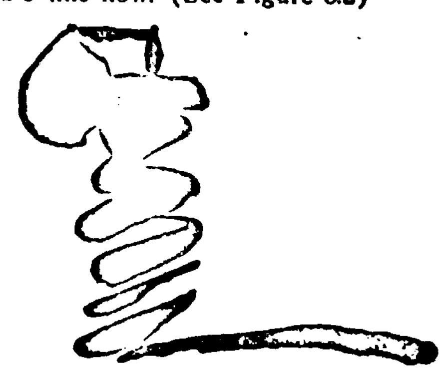


FIGURE 8.2
SAMPLE ITEM OF AMBIGUOUS INK BLOTS, CI317A

[13] Developed at Psychological Research Unit No. 2. Chief contributor: Staff/Sgt. Arthur Z. Cerf.

For this sample problem, the listed responses are: A—Distilling apparatus; B—Baseball catcher's mitt; C—Coiled spring; D—Fire-hose nozzle; E—Bar stool; F—Typewriter; G—Oriental lantern; H—Two-wheeled cart; I—Stairway; J—Desk calendar; K—Country mail box; L—Crown; M—Turtle; N—Cat's head; and O—Smoker's pipe.

In the sample problem which you have just finished, almost everyone is able to find: A. Distilling apparatus; C. Coiled spring; E. Bar stool; and O. Smoker's pipe. Some people are also able to find: G. Oriental lantern; K. Country mail box; and N. Cat's head. A few are able to find: I. Stairway; and L. Crown. Practically no one is able to find B. Baseball catcher's mitt; D. Fire-hose nozzle; F. Typewriter; H. Two-wheeled cart; J. Desk calendar; and M. Turtle. There are no absolutely right or wrong answers. The important thing is that you indicate accurately the objects you can see.

(2) *Scoring.*—The scoring formula for this test is the number of responses. This assumes that the quality of response is irrelevant.

Ambiguous Figures, CI316A [14]

This test is believed to measure the ability of the examinee to evoke as many relationships as possible from a pair of geometric figures.

Description. (1) *Administration.*—The following are the directions for the test:

This is a test of your ability to find relationships between geometric figures. Look at sample problem No. 1 below. (See figure 8.3.)

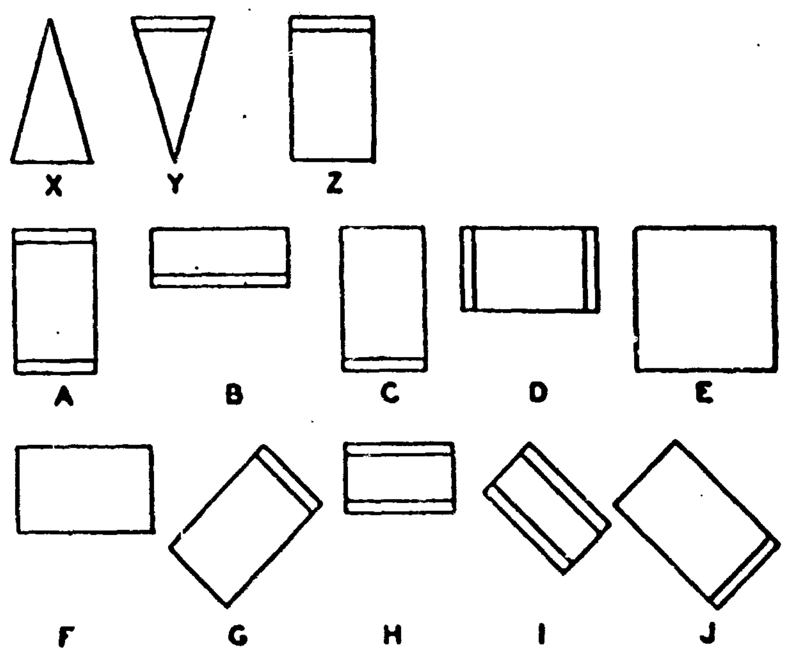


FIGURE 8.3
SAMPLE ITEM OF AMBIGUOUS FIGURES, CI316A

[14] Developed at Psychological Research Unit No. 2. Chief contributors: Tech./Sgt Paul C. Davis, Capt. John I. Lacey, Jeanne L. Lipman, Tech./Sgt. Gerald H. Shirley.

X is related to Y in several ways:

1. Both figures have the same shape.
2. Y is X rotated 180°.
3. An additional line has been drawn from Y.

Your task is to find which of the choices, A through J, bear 2 or more of the same relationships to Z as Y bears to X.

Examine choice A. It fulfills relationship 1, since both figures have the same shape. It has also been rotated 180° fulfilling relationship 2. Relationship 3 is satisfied, as a line has been drawn across the top part of the figures. Since choice A fulfills the requirement of having at least 2 relationships to Z, it is correct. Blacken A opposite No. 1 on your answer sheet.

Now examine choice B. It does not fulfill relationship 1, because both figures do not have the same shape. Since the line through Z is at the top, while the line in B is at the bottom, B may be considered to be rotated 180°. It does not fulfill relationship 3, because an additional line has not been drawn across the top of the figure. Since it meets only 1 relationship, it cannot be considered a correct choice. *Notice that size is not to be considered as one of the relationships.*

Look at choice C. It fulfills relationship 1, because it has the same shape as Z. Since it has beeen rotated 180°, it fulfills relationship 2. It does not have relationship 3. However, since it bears at least two correct relationships, it is a correct choice. Blacken C opposite No. 1 on your answer sheet now.

Go through the remaining choices, D through J, and blacken the correct answers. Do this now. You should have marked choices A, C, D, and H. If you have not marked them correctly, do so now.

The time limits are as follows: Part I—12 minutes, part II—12 minutes.

(2) *Scoring.*—The scoring formula for this test is the number right.

Evaluation of Fluency Tests

Seven tests of fluency were constructed. Complete coverage of the area was not obtained, inasmuch as no tests of fluency based upon numbers or pictures were constructed, because of difficulties involved in constructing tests using these media.

Four tests (Camouflaged Words, CI323A; Verbal Recognition, CI322A; Verbal Recognition, CI322B, and Ambiguous Figures, CI316A) were administered to samples ranging from 400 to 2,500 for correlational and factor-analysis purposes along with a large group of experimental and classification tests. The intercorrelational data were not available at the time this was written.

FACTOR ANALYSIS OF JUDGMENT TESTS

The Data

In order to analyze types of judgment items systematically, a special judgment-and-reasoning test was constructed by the Psychological Branch of the Office of the Air Surgeon. A factor analysis utilizing this test was later performed at Psychological Research Unit No. 3.[18]

A subjective analysis of judgment items had suggested that the following elements were involved in answering them:

[18] Chief contributors: Staff/Sgt. Benjamin Fruchter, Capt. Lloyd G. Humphreys.

1. Word knowledge.
2. Factual information of a practical type.
3. Logical reasoning ability.
4. Mechanical comprehension and information.
5. Ability to make common-sense judgments.

The Tests.—To measure these five elements, several tests were constructed, including four separate tests of reasoning ability and two tests of mechanical comprehension. In addition, other variables whose relationships with judgment items could be expected to throw light on the psychological make-up of judgment items were prepared for inclusion in a judgment-and-reasoning test. The following is an outline of the test:

Variable 1, general vocabulary, item Nos. 1-10: These items are intended to provide a measure of word knowledge. A sample item follows:

Deft:
- A. Skillful.
- B. Insane.
- C. Clumsy.
- D. Split.
- E. Light.

Variable 2, ten most valid judgment items in the AAF Qualifying Examination, AC10A, items 11-20: This variable includes items, each one of which had been found to have positive correlation with graduation-elimination from pilot training. A sample item follows:

A radio aerial has broken. It formerly led from the roof of a house to the top of a 30-foot pole set in the ground outside. It is not safe for a man to climb this pole as it is only 4 inches in diameter at the base and 2 inches at the top. Of the following, the most practical way to put up a new aerial would be to
- A. Use a 30-foot ladder.
- B. Take the pole down, attach the aerial, and then reset the pole.
- C. Use a fishing pole to hook the aerial to the top of the pole.
- D. Make a noose in the end of the aerial and throw it over the top of the pole.
- E. Build a light scaffold around the pole.

Variable 3, commonsense judgment, items 21-30: This part is intended to measure ability to make commonsense judgments rather than logical reasoning ability or mechanical comprehension and information. A sample item follows:

A bomber squadron is over enemy territory on its way to bomb an oil refinery when one of the observers notices an advanced enemy airdrome. He notifies the squadron leader. It would be best for the squadron leader to
- A. Order the bombers in the squadron to continue as planned and report the location of the enemy airdrome on arrival at their base.
- B. Order the squadron to circle the airdrome while he radios its position to his base.
- C. Order half the bombers in the squadron to bomb the enemy airdrome and the other half to carry out the mission against the oil refinery.
- D. Order all the planes in the squadron to bomb the enemy airdrome and return to their base.

E. Order the squadron to continue as planned while he returns to his base to report the location of the enemy airdrome.

Variable 4, mechanical judgment, items 31-40: This part is intended to measure ability to make judgments based mainly on mechanical comprehension and information. Following is a sample item:

A soldier accidentally bent the front sight of his rifle somewhat to the right. Until the sight is repaired, the gun will shoot too far to the

A. Right and too high.
B. Left and too high.
C. Left, but neither too high nor too low.
D. Right and too low.
E. Right, but neither too high nor too low.

Variable 5, logical-reasoning judgment, items 41-50: This variable is intended to measure chiefly the ability to make judgments almost wholly on the basis of logical reasoning. Common sense and mechanical comprehension are judged to be unimportant. A sample item follows:

An officer is in command of an advance unit in a night foray against distant enemy lines to test their strength in preparation for an assault against them. His orders are to send up a red flare if the enemy is not prepared for an attack, a blue flare if the enemy is prepared for an attack, a red flare and then a blue flare if the enemy is preparing an attack of its own, and a blue flare and then a red flare if the enemy is already beginning an attack. The officer finds the enemy beginning an attack, but instead of sending up first a blue flare and then a red one, by mistake he sends up a red flare first. If he realizes his mistake immediately, it would be best for him to

A. Wait several minutes and then send up a blue flare with a red flare following immediately after it.
B. Send up a blue flare right away and then a red flare immediately after it.
C. Send up a blue flare, wait a minute, and then send up a blue flare with a red flare immediately following it.
D. Send up a blue flare, wait a minute, and then send up a red flare with a blue flare immediately following it.
E. Send a man back to the base to report the mistake.

Variable 6, deductive reasoning, items 51-60: This test is designed to measure ability to draw a logical conclusion from a problem situation. Following is a sample item:

An inspector general has an appointment in a city one hundred miles away. If the train on which he must travel is late, he will miss his appointment. If the train is not late, he will miss the train. We do not know whether the train is late. With this information, we can state positively that

A. He will not be able to keep his appointment.
B. He will be able to keep his appointment.
C. There is no way of telling whether he will be able to keep his appointment.
D. He will have to take a later train.
E. He will have to wait for the train.

Variable 7, arithmetic reasoning, items 61-70: This variable measures arithmetical reasoning ability. Numerical computation is minimized. A sample item follows:

An Army truck goes 10 miles on a gallon of gasoline and 60 miles on a quart of oil. If there were 8 gallons of gasoline in the tank and 1½ gallons of oil in the motor, how far could this truck go?

A. 70 miles.
B. 80 miles.
C. 90 miles.
D. 170 miles.
E. 440 miles.

Variable 8, information in judgment, items 71-83: Thirteen items were included in this part to test knowledge that would be required to answer the ten judgment items included in variable 2. A sample of the items in variable 8 is the following item, which was constructed to test for information considered crucial in answering item 14 in variable 2.

A wooden flagpole 30 feet tall is to be erected in front of a school-house. No guy wires or other supports can be used. Of the following methods of setting the pole in place, the one that would be easiest and yet adequately safe would be to

A. Dig a hole 2-feet deep, place the pole upright in it, and replace the soil in the hole, tamping it solid.
B. Dig a hole 4-feet square and 4-feet deep, mix enough concrete to fill the hole, place the pole upright in the concrete, and support it in that position until the concrete hardens.
C. Secure a large block of granite, drill a hole in it large enough for the pole to fit in, place the block in a hole of appropriate size, and slide the pole into the hole in the block.
D. Bore a hole about 2-feet deep with an auger slightly larger in diameter than the pole, insert the pole and fill in around it with sand.
E. Dig a hole 2-feet square and 5-feet deep, place the pole upright in it, and fill in around the pole with coarse gravel.

Variable 9, mechanical comprehension, items 84-93: This variable includes 10 mechanical comprehension items similar to those in test CI903A (see page 304 for sample item).

Variable 10, reasoning in reading, items 94-103: This variable consits of 10 reading comprehension items selected because they appear to measure a component of reading comprehension called reasoning in reading (2). Ability to make inferences is stressed. A sample item follows:

One of the most beautiful military replies I've ever heard of was given in India by a captain who had lost a steam roller. The Government sent him several forms to be filled out before it could be replaced. On one form was the question: "Reason for loss?" The captain filled in the words: "Eaten by white ants." He never heard another word about it, but in due course of time his replacement arrived.

It is most probable that the captain:

A. Did not really know what happened to his steam roller.
B. Told the truth about the steam roller.
C. Was disgusted at having to fill out so many forms.
D. Did not dare tell what had really happened to the steam roller.
E. Did not care whether his steam roller was replaced.

Variable 11, syllogisms, items 104-113: This part is intended to measure logical reasoning ability unaffected by habitual modes of thought. The directions are:

Read each one of the following items as you come to it. Then decide whether the last sentence in each item necessarily follows if the preceding statements in the item are accepted as true. If you think that the last sentence is a necessary conclusion, make a mark in the corresponding space on your answer sheet lettered A. If you think that the last sentence is not a necessary conclusion, make a mark in the corresponding space on your answer sheet lettered B.

A sample item is: "Only thieves hide jewels. This man hid jewels. Therefore, he must be a thief."

Variable 12, mechanical movements, items 114-123: This variable includes 10 mechanical movements items adapted (by permission) from Thurstone's mechanical movements test. The items are similar to the sample given on page 317.

Variable 13, figure analogies, items 124-136: This part contains 13 figure analogies items taken (by permission) from the nonverbal reasoning test of the 1942 National Teacher Examinations. They are similar to the customary items of this type (see page 105).

Variable 14, pattern reasoning, items 137-150: The last variable consists of 14 pattern-analogies items taken (by permission) from the nonverbal reasoning test of the 1942 National Teacher Examinations. A sample item is shown in figure 8.4. The directions are:

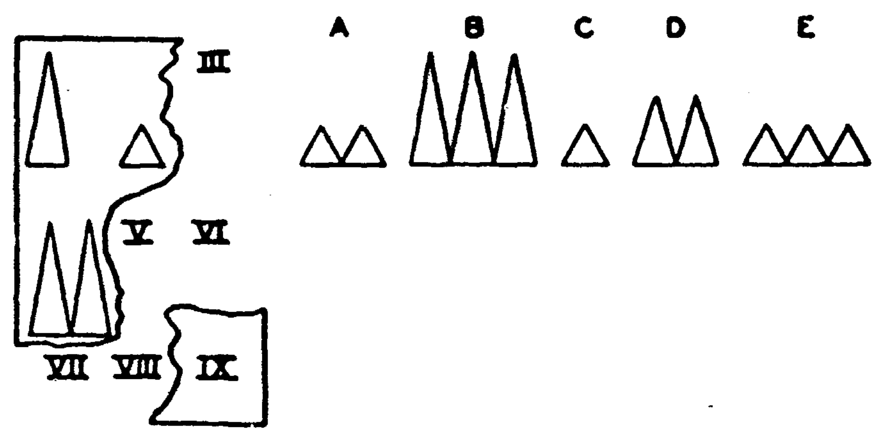


FIGURE 8.4
SAMPLE ITEM OF PATTERN-REASONING SECTION OF JUDGMENT AND REASONING TEST

Each item of the following section consists of nine diagrams arranged in rows of three each. The diagrams form a pattern. Some of the nine diagrams are omitted, and the problem is to determine which of the five figures given as choices belongs in the ninth space (the third space in the third row).

In assembling each part of the judgment-and-reasoning battery every effort was made to select items that did not overlap the mental functions supposed to be tested by other parts of the test. Data that were available regarding the difficulty of items made it possible to use items having a sufficiently wide spread of difficulty to make each part approach a miniature power test. It would have been preferred to use a larger number of items in each part to provide greater reliability of measurement, but limitations of time and the desirability of confining the scoring to one side of a standard answer sheet prevented the use of longer subtests. To permit virtually every examinee to attempt every item, the present form of the battery requires 3 hours for administration.

The Samples

The Judgment and Reasoning Test was administered to a sample of 689 eleventh- and twelfth-grade boys in the Stuyvesant High School, New York City [16] and to a sample of 1,024 aviation students classified for pilot training at Psychological Research Unit No. 2.

Tables 8.9 and 8.10 present the intercorrelations of the 14 part-scores of the Judgment and Reasoning Test for the high school and aviation-student samples, respectively. Table 8.11 gives the centroid loadings and communalities for both analyses. Table 8.12 gives the rotated factor loadings for both analyses.

The Factors

In the following paragraphs the rotated factor loadings from the two analyses will be discussed together, since the six factors are practically identical in both. The analysis based on high school students is labeled I, that on aviation students, II.

Rotated factor I is defined by the following data:

Test number	Test name	Analysis *I*	Analysis *II*
5	Logical reasoning judgment	0.41	0.41
7	Arithmetic reasoning	.40	.38
13	Figure analogies	.34	.02
4	Mechanical judgment	.26	.03
14	Pattern reasoning	.24	.38
3	Commonsense judgment	.23	.01
6	Deductive reasoning	.19	.35

This factor is undoubtedly the reasoning I (general-reasoning) factor usually best defined by an arithmetic-reasoning test. Some of the discrepancies in the loadings are due to differences in variability in the two groups. On most reasoning tests, for example, the aviation-student population seems to be more variable, as can be seen from a comparison of the variances of tests loaded with this factor. The difference in loadings for figure analogies is sufficiently large and in the opposite direction from that predictable from the variances, however, to suggest a difference in function tested in the two groups by this test. This test differs

[16] Indebtedness is hereby expressed for the cooperation of the New York school authorities which made this administration possible.

TABLE 8.9.—*Correlation matrix of the Judgment and Reasoning Tests based on a sample of eleventh- and twelfth-grade high-school boys (N=689)*[1]

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 Vocabulary	..	20	15	20	21	17	28	14	25	31	23	06	20	21
2 Judgment AC10A	20	..	23	39	18	15	29	30	45	19	14	25	21	25
3 Commonsense (pure) judgment	15	23	..	19	29	16	28	18	27	24	22	13	20	22
4 Mechanical judgment	20	39	19	..	24	20	32	33	52	19	19	28	25	26
5 Logical reasoning judgment	21	18	29	24	..	17	34	19	21	21	18	13	26	24
6 Deductive reasoning	17	15	16	20	17	..	24	10	20	20	19	10	18	15
7 Arithmetical reasoning	28	29	28	32	34	24	..	25	33	26	30	13	35	28
8 Information in judgment	14	30	18	33	19	10	25	..	31	20	17	20	17	24
9 Mechanical comprehension	25	45	27	52	21	20	33	31	..	23	26	37	33	35
10 Reasoning in reading	31	19	24	19	21	20	26	20	23	..	26	12	24	30
11 Syllogisms	23	14	22	19	18	19	30	17	26	26	..	14	28	25
12 Mechanical movements	06	25	13	28	13	10	13	20	37	12	14	..	25	20
13 Figure analogies	20	21	20	25	26	18	35	17	33	24	28	25	..	46
14 Pattern reasoning	21	25	22	26	24	15	28	24	35	30	25	20	46	..

[1] Decimal points omitted.

TABLE 8.10.—*Correlation matrix of the Judgment and Reasoning Tests based on a sample of unclassified aviation students (N=1,024)*[1]

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 Vocabulary	..	19	27	22	25	26	27	18	09	43	23	14	24	20
2 Judgment AC10A	19	..	25	24	16	14	21	20	25	21	15	17	20	17
3 Commonsense (pure) judgment	27	25	..	20	26	20	20	14	17	24	16	13	21	15
4 Mechanical judgment	22	24	20	..	18	21	23	23	30	23	19	24	21	20
5 Logical reasoning judgment	25	16	26	18	..	34	43	19	24	28	18	22	29	29
6 Deductive reasoning	26	14	20	21	34	..	38	16	22	28	15	20	21	25
7 Arithmetical reasoning	27	21	20	23	43	38	..	23	33	30	21	27	13	37
8 Information in judgment	18	20	14	23	19	16	23	..	19	17	11	17	16	23
9 Mechanical comprehension	09	25	17	30	24	22	33	19	..	17	14	36	21	29
10 Reasoning in reading	43	21	24	23	28	28	30	17	17	..	24	16	33	29
11 Syllogisms	23	15	16	19	18	15	21	11	14	24	..	15	25	21
12 Mechanical movements	14	17	13	24	22	20	27	17	36	16	15	..	25	23
13 Figure analogies	24	20	21	21	29	21	13	16	21	33	25	25	..	50
14 Pattern reasoning	20	17	15	20	29	25	37	23	29	29	21	23	50	..

[1] Decimal points omitted.

TABLE 8.11.—*Centroid factor loadings and communalities for the Judgment and Reasoning Tests*[1]

Test	High-school boys (N=629)							Unclassified aviation students (N=1,024)						
	I	II	III	IV	V	VI	h^2	I	II	III	IV	V	VI	h^2
1 Vocabulary	42	18	12	−15	14	−11	28	50	−24	−25	−20	−09	−09	43
2 Judgment AC10A	53	−29	11	−04	−08	−08	39	41	04	−20	13	09	12	25
3 Commonsense (pure) judgment	44	11	09	08	−09	18	26	42	−05	−20	−08	23	14	30
4 Mechanical judgment	59	−33	07	10	07	−11	49	47	12	−20	18	−09	−09	32
5 Logical reasoning judgment	46	18	15	23	−14	09	35	55	06	16	−27	11	15	44
6 Deductive reasoning	36	11	07	09	17	−04	19	50	09	08	−24	−05	−04	33
7 Arithmetical reasoning	58	16	14	11	−08	−11	41	59	27	08	−26	−23	07	55
8 Information in judgment	45	−19	10	−15	−14	10	30	38	10	−10	10	−11	13	20
9 Mechanical comprehension	65	−32	−09	03	11	−09	57	49	32	07	20	09	−08	40
10 Reasoning in reading	47	22	07	−24	10	07	35	55	−27	−10	−15	−05	−09	42
11 Syllogisms	45	20	−06	−03	10	06	26	39	−15	−04	03	−03	−03	18
12 Mechanical movements	40	−27	−24	10	07	18	34	45	22	11	15	09	−19	33
13 Figure analogies	55	18	−36	10	−14	−15	52	54	−34	25	27	09	04	55
14 Pattern reasoning	56	13	−27	−12	−17	−06	45	57	−13	36	15	−16	22	57

[1] Decimal points omitted.

783120—47—11

Table 8.12.—*Rotated factor loadings and matched-half reliability coefficients for the Judgment and Reasoning Tests*[1]

Variable	Aviation students (pilot) (N=1,026)[2]								High-school boys (N=689)							
	I	II	III	IV	V	VI	h^2	${}^{a}r_{1I}$	I	II	III	IV	V	VI	h^2	${}^{a}r_{1I}$
1 Vocabulary	07	05	10	52	05	−01	29	49	03	05	10	62	06	10	41	70
2 Judgment AC10A	17	13	51	22	15	04	38	46	00	13	34	19	12	25	24	34
3 Commonsense (pure) judgment	23	10	13	26	03	32	25	23	01	07	15	28	12	40	28	20
4 Mechanical judgment	26	14	47	23	35	00	48	42	03	10	43	27	23	00	32	27
5 Logical reasoning judgment	41	06	09	28	00	27	33	26	41	15	00	23	23	32	42	39
6 Deductive reasoning	19	04	03	33	19	04	18	35	35	05	06	34	25	10	32	55
7 Arithmetical reasoning	40	14	19	42	03	10	41	59	58	01	23	33	22	06	54	69
8 Information in judgment	−04	12	44	22	02	20	30	36	16	12	35	17	05	09	20	36
9 Mechanical comprehension	17	32	44	28	41	01	57	69	18	18	35	03	46	08	40	57
10 Reasoning in reading	−03	12	10	55	03	17	36	50	07	19	04	58	12	10	40	63
11 Syllogisms	11	20	00	42	12	16	26	28	03	21	12	32	09	06	17	27
12 Mechanical movements	07	31	22	00	37	22	34	65	12	18	22	09	47	00	32	61
13 Figure analogies	34	51	01	38	02	06	53	64	02	66	05	26	15	11	54	76
14 Pattern reasoning	24	61	06	04	11	08	45	66	38	59	15	20	04	05	56	72

[1] Decimal points omitted.
[2] Obtained by correcting reliability coefficients based on sample of 1026 aviation students for differences in range of talent.
[3] Obtained by correlating matched halves of each test and correcting by means of the Spearman-Brown formula.

from Figure Analogies, CI212AX1, in that the principles underlying the analogies are generally more subtle and difficult. In these respects the test is probably more similar to Figure Classification, CI213AX1, also known to have a near zero loading on the general-reasoning factor in an aviation-student sample. It will be noticed that high-school students also tend to solve mechanical-judgment and commonsense-judgment items by reasoning.

When test items are too difficult for the group examined to handle in the intended fashion, other abilities may be called upon if motivation is high. The easiest thing to do under these circumstances is to seek a solution in the misleads rather than to seek the alternate which best fits the reasoned solution. This may involve a different sort of reasoning than that called for by an arithmetic-reasoning test.

Rotated factor II is defined by the following data:

Test number	Test name	Analysis *I*	Analysis *II*
14	Pattern Reasoning	0.61	0.59
13	Figure Analogies	.51	.66
9	Mechanical Comprehension	.32	.18
12	Mechanical Movements	.31	.18

This factor was not well-defined. Additional information is now available, however, which explains the difficulty in naming it. Tests of figure analogies have been found with substantial loadings on three factors other than the reasoning I factor. One of these has been termed integration III, another reasoning II, and the third, reasoning III (see pp. 119f.). This factor is probably closest akin to reasoning II, but the loading of figure analogies is much larger than usual in any factor. It is possible that there is a combination of two factors here; consequently, rotated factor II will be named reasoning II only with considerable hesitation. The Pattern Reasoning Test has never appeared in any other battery for analysis; nor has this form of Figure Analogies. It may be that these two forms have by some fortunate circumstance achieved an unusually high degree of purity (as reasoning tests go) for one of the reasoning factors.

Rotated factor III is defined by the following data:

Test number	Test name	Analysis *I*	Analysis *II*
2	Judgment, AC10A	0.51	0.34
4	Mechanical Judgment	.47	.43
8	Information in Judgment	.44	.35
9	Mechanical Comprehension	.44	.18
12	Mechanical Movements	.22	.22

This is clearly the mechanical-experience factor. Practically all of the loadings show a drop from the high-school to the aviation-student sample. This is explained by the decreased variability of scores of the latter group on mechanical tests, which can be traced to the mode of their selection. It was early recognized that the judgment test in AC10A had a heavy mechanical variance. The selection of items by correlation with the pilot criterion—also heavily weighted with the mechanical factor—also favored this state of affairs in later judgment tests.

Rotated factor IV is defined by the following data:

Test number	Test name	Analysis I	Analysis II
10	Reasoning in Reading	0.55	0.58
1	Vocabulary	.52	.62
7	Arithmetic Reasoning	.42	.33
11	Syllogisms	.44	.32
13	Figure Analogies	.38	.26
6	Deductive Reasoning	.33	.34
5	Logical Reasoning Judgment	.28	.28
9	Mechanical Comprehensions	.28	.03
4	Mechanical Judgment	.23	.27
3	Commonsense Judgment	.26	.28

This, the verbal factor, seems to be about equally well-defined in the two groups. The larger variance of the Vocabulary test in the aviation-student sample has increased its loading with the factor slightly. Both analyses show that judgment tests tend to have low but probably significant loadings in the verbal factor. It would be desirable to depress this variance still further, particularly if a judgment test is to be used for pilot selection.

Rotated factor V is defined by the following data:

Test number	Test name	Analysis I	Analysis II
9	Mechanical Comprehension	0.41	0.46
12	Mechanical Movements	.37	.47
4	Mechanical Judgment	.35	.23
6	Deductive Reasoning	.19	.25

The visualization factor is about equally clear-cut in the two analyses. The Mechanical Judgment test is a poorer measure of the visualization factor in the aviation-student group, members of which have more mechanical information. Since it is impossible to state the statistical significance of a difference in two factor loadings, however, one can only speculate about a difference of this size.

Rotated factor VI is defined by the following data:

Test number	Test name	Analysis I	Analysis II
3	Commonsense Judgment	0.32	0.40
5	Logical Reasoning Judgment	.27	.32
2	Judgment AC10A	.04	.25
12	Mechanical Movements	.22	.00
8	Information in Judgment	.20	.09

The high-school sample loadings with this factor (which had been identified first in the analysis based upon aviation students) represent the nearest approach to those for the aviation-student sample that could be achieved in the rotations. Without knowledge of the existence of the factor in the aviation-student group it would have been very easy to have treated the sixth factor in the high school group as a residual. Whatever the nature of factor VI, it is clear that any unexplained validity of variable 3 (Commonsense Judgment) is probably due to its loading here. Since the communality of this variable is actually somewhat greater than its estimated reliability, we can conclude that its non-chance variance is completely accounted for by the common factors in these analyses.

In naming this factor, the fact that it is better defined in the aviation-student group is of considerable importance. The variances of variable 3, Commonsense Judgment, are almost identical in the two samples, but there is considerable difference in the mean scores. If factor VI were an interest factor, it would be relatively easy to account for such a shift. The interest could be in things aviation or military. The content of the test which best defined the factor is congruent with this hypothesis. It would therefore be reasonable to assume that this interest factor would be stronger in the aviation-student than in the high school group. A second hypothesis is that the factor represents judgment, an ability traditionally unrelated to academic work. This ability might also be at a higher average level in the more mature, less academic aviation-student group.

It is very reasonable to identify this factor with the one called judgment in the foresight-and-planning analyses (see ch. 9), which was best defined by a set of work-planning judgment items. Other tests appearing with the factor in these analyses were the Practical Estimations Test, Sequence of Maneuvers, and Competitive Planning. Judgment is certainly a more plausible designation than interest, for this group of tests.

Validities of the Judgment and Reasoning Tests

Scores on the 14 parts of the Judgment and Reasoning Test were correlated with graduation-elimination in elementary pilot training. The

TABLE 8.13.—*Validation data for the fourteen parts of the judgment and reasoning battery based on the graduation or elimination of 746 pilots in primary training (p_s=0.86)*

Part	Type of item	SD_t	M_g	M_e	r_{bis}	$_c r_{bis}$ [1]
1	Vocabulary	2.15	7.03	7.24	−0.05	0.00
2	Judgment AC10A	1.54	6.62	6.43	.07	.15
3	Commonsense (pure) Judgment	1.57	7.34	6.98	.12	.17
4	Mechanical Judgment	1.64	7.08	6.87	.07	.16
5	Logical Reasoning Judgment .	1.84	5.05	4.76	.09	.15
6	Deductive Reasoning	1.79	6.52	6.21	.10	.16
7	Arithmetical Reasoning	2.38	5.32	5.01	.07	.16
8	Information in Judgment	1.83	8.27	8.29	.00	.07
9	Mechanical Comprehension ...	1.70	7.85	7.17	.22	.32
10	Reasoning in Reading	1.89	7.56	7.28	.08	.12
11	Syllogisms	1.66	5.94	5.82	.04	.08
12	Mechanical Movements	2.01	6.40	6.36	.01	.14
13	Figure Analogies	3.00	8.01	7.36	.12	.20
14	Pattern Reasoning	2.94	6.10	5.42	.13	.22

[1] Assuming an unrestricted standard deviation of 1.83.

validation data are presented in table 8.13, for a sample tested at Psychological Research Unit No. 2 with the November 1943 classification battery.

Conclusions

Validities obtained for variable 3 (Commonsense Judgment) indicate that the judgment factor has an appreciable degree of pilot validity. This test variable has only one other sizeable loading in the aviation-student analysis, and that is on the verbal factor, which has no positive pilot validity.

Shifts in factor patterns between the high-school population and the aviation-student population are of at least two sorts. The one has a very simple explanation—differences in range of talent affect factor loadings the same as they affect other correlation coefficients. Although there are a few tests that are exceptions, the verbal and general-reasoning factors are better defined, and the mechanical-experience factor is more poorly defined in the aviation-student sample for this reason.

There is some evidence, on the other hand, that certain tests measure different abilities in the two populations. The Figure Analogies test shows a shift from the reasoning I factor in the high-school group to the reasoning II factor in the aviation-student group. The judgment factor is more clear-cut and is defined by higher loadings in the aviation-student group. Other differences between the two analyses, unexplained by the differences in variability, are not as large and, on the basis of present data, probably cannot be distinguished from sampling fluctuations.

SUMMARY AND EVALUATION OF JUDGMENT TESTS

Judgment was found to be the most frequently mentioned psychological category to which flight instructors referred when giving reasons for eliminating pilots. In an attempt to measure and better define this category, a series of practical judgment tests was constructed. Factor analysis revealed a judgment factor, best defined by the work-planning type of item.

In an attempt to understand the informational basis of judgment, a series of practical-estimation tests was constructed. Items calling for absolute estimates of time, distance, etc., were found to be uncorrelated with other practical-estimation and practical-judgment items. Items calling for relative estimates had satisfactory communality with judgment items. Of this latter type, those items calling for relatively complicated estimates involving time as well as distance and size were found to be significantly loaded with the judgment factor, whereas those items calling for relatively simple estimates of size and speed contained no judgment loading. One inference is that the judgment factor is a thinking, rather than a perceptual or memory ability.

Another attempt to explore the judgment category was based upon the assumption that the fluency with which hypotheses can be evoked would be an important element in arriving at the correct solution to a judgment problem situation. A series of tests of fluency was constructed. Analytical results for these tests were not available at the time this volume was being written.

BIBLIOGRAPHY

(1) *An Historical Introduction to Aviation Psychology.* National Research Council, 1942.

(2) Davis, F. B. Two new measures of reading ability, *J. Educ. Psychology,* May 1942. 33, 365–372.

(3) Kelly, E. L. *The Development of a Scale for Rating Pilot Competency,* Civil Aeronautics Administration, Washington, D.C., 1943.

(4) Thurstone, L. L. *Primary Mental Abilities,* University of Chicago Press, Chicago, Ill., 1938.

CHAPTER NINE

Foresight and Planning Tests[1]

INTRODUCTION

One feature contributing to success as a pilot is the ability to plan a series of maneuvers or activities and to foresee and avoid difficulties that may arise in their execution. Such is the statement of the problem in common-sense terms.

Job-Analysis Information

Job-analysis studies give foresight and planning a high-ranking position among pilot qualifications.

In elementary training.—In an analysis of the faculty-board proceedings[2] which report why student pilots were eliminated from further training at the elementary stage, lack of foresight and of planning ability was reported in over a third of the cases.[3] Instructors stated that the students lacking this ability failed to plan ahead properly for landings, made incorrect and dangerous entrances and exits from traffic, were unable to plan forced landings properly, flew the traffic pattern improperly, etc.

In another analysis of faculty-board proceedings, data were fractionated according to the number of flying hours and the results in table 9.1 were obtained. The frequency of deficiency in foresight and planning among eliminees at different stages of primary pilot training is thereby shown. From this it can be seen that the deficiency remains uniformly important after the first five hours of flying lessons.

TABLE 9.1.—*Percentage of eliminees from pilot training showing deficiencies in foresight and planning at different stages in training*

Hours of elementary training	Percent of eliminees deficient
1–4	3
5–8	43
9–12	44
13–16	44
17–24	42
25–36	43
36–60	36

In another study in which rating scales were filled out by instructors for 1,303 cadets, lack of foresight and planning was named as a cause of

[1] Written by T/Sgt. Sanford J. Mock and the editor.

[2] In aviation reports foresight and planning are usually regarded as one ability. It remains to be demonstrated that this is a psychological fact, or even that the two are irreducible categories.

[3] See table 1.5.

elimination in 43 percent of the cases. This placed foresight and planning as the second most frequently indicated deficiency in a group of 20 categories.

In a study in which instructors rated students after 8 to 10 hours of flying instruction, biserial coefficients of correlation were computed between various traits and graduation elimination from primary training. The rating on foresight and planning for 369 cadets correlated 0.67 with graduation elimination. In the same study, foresight and planning was mentioned as a deficiency in 38 percent of the eliminees.

In a study of landing planes, it was reported that while flying the downwind leg, the pilot must plan where to make the 90° turn into the base leg. Before the turn toward the landing lane, the throttle is cut and a glide established. The pilot must judge accurately where to cut the throttle, how fast the airplane is gliding, and when to make the gliding turn. Placing the gliding turn requires accurate judgment and planning.

Eighty-eight students in primary training filled out forms indicating their greatest worries during landing. Placing of the gliding turn was the fourth most frequently mentioned worry. Eighty-four students and their instructors were interviewed individually to determine what they thought were the chief problems in landing. Placing the gliding turn correctly was the third most frequently mentioned item.

Later training and combat.—Similar facts were obtained at advanced stages of training. A summary was made of the frequency with which various reasons were stated by the faculty board for the elimination of 100 students in single-engine training and 100 students in twin-engine training. Forty-seven percent of the eliminees from single-engine training and 23 percent of the eliminees from twin-engine training were listed as deficient in foresight and planning.[a]

Data are available on the final disposition of 100 unsatisfactory pilots reclassified by a flying evaluation board in operational training or in combat. Twenty-two percent of the 100 reclassified pilots were listed as deficient in the category of intelligence and judgment of which foresight and planning was regarded as a component. This deficiency was given as a reason for reclassification.

Through the Informational Intelligence Division of the Army Air Forces, reports were obtained of interviews with American, British, and Chinese individuals and groups in combat, concerning efficiency of air crew, morale, training, operations, etc. The interview material was organized into psychological and quasi-psychological concepts relevant to aircrew selection.

The statements about the fighter pilot include:

Automaticity in combat or while in flight. A good pilot is busy all the time—must plan ahead* * * You must plan what you are going to do while on the ground —you must think in advance what you are going to do up in the air.* * *foresight and planning are important for the bomber pilot * * * It is resource,

[a] See table 1.5.

not daring, that makes a successful operation. The more time spent in preparing the flight, the better the chances of success.

The Informational Intelligence Division study also reported that a characteristic of the successful navigator is "planning and foresight, including being prepared and fully briefed, convinced of what he has to do and what to do in an emergency."

TESTS OF PATHWAY PLANNING

It is often necessary for the pilot or navigator to plan an aerial route, subject to certain restrictions. Finding the target and returning to the home base are examples of situations in which pathway planning is required. While the target, or base, might be reached by several different routes, the limitations of the situation may actually permit only one approach. The gasoline load, position of enemy antiaircraft, likelihood of meeting enemy fighters, weather conditions—these are types of limitations which may force the pilot or navigator to select the one, and only one, appropriate path to the objective. The need for the ability to plan routes prompted the development of the Route Planning and Planning a Circuit tests.

Route Planning, CI411AX [a]

Route Planning and Map Planning were constructed as paper-and-pencil forms of the Foresight and Planning Maze Test, CI405A, an apparatus test. CI405A consists of a slot-maze board to be used with a stylus. The parallel straight alleys intersect at acute angles forming diamond-shaped islands, each with an electric-light bulb in its center. One of the bulbs is lighted to become the goal of the moment. Various paths lead to the goal, some being short and economical, and many others are longer and less direct. The blocked passages are visible to the examinee who, on the signal, inserts the stylus at the entrance to the maze. A light appears on one of the diamonds and remains lighted for 15 seconds during which the examinee plans his course but does not move. When the light goes out, the examinee immediately starts for the goal diamond and is allowed 10 seconds to reach it. This cycle of events is then repeated with a new starting point and a new goal.

Description.—In Route Planning, CI411AX, the examinee must plan a path successively from four points on the periphery of a printed maze to a goal box in its center (see fig. 9.1). There are four item numbers, one at each corner of the maze. Each number is the starting point for an item. The darkened square near the center of the maze is the common goal. The task is to locate the one point through which one must pass in going from each starting point to the goal. Each group of four items is based on a pair of identical maze patterns, one that the examinee studies briefly and one that he uses in making his answers. In the latter, letters mark the various pathways to the center.

[a] Developed at Psychological Research Unit No. 3. Chief contributor: Lt. William M. Wheeler.

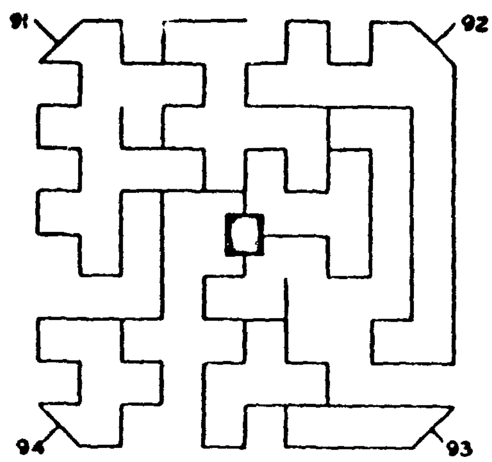


FIGURE 9.1
SAMPLE STUDY-MAZE OF ROUTE PLANNING, CI411AX

(1) *Internal characteristics.*—The test consists of 2 sample study mazes, providing 8 recorded but unscored sample items, 4 to each maze, and 36 scored items based on 9 different maze patterns. There are 2 parts of 16 and 20 items respectively. The mazes in part I have fewer lines and are simpler than those in part II. Each maze in the test proper covers an area 10 inches by 6½ inches. The first sample maze is 2½ inches square and the second sample maze is 4½ inches square.

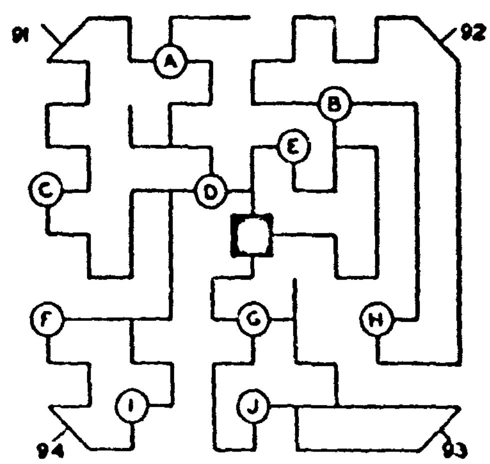


FIGURE 9.2
SAMPLE ANSWER-MAZE OF ROUTE PLANNING, CI411AX

(2) *Administration.*—The amount of time allowed for studying each diagram and for answering questions on its mate varies as follows:

Item No.	Study time (minutes)	Answer time (minutes)	Item No.	Study time (minutes)	Answer time (minutes)
1-4	1.50	0.75	21-24	1.50	1.00
5-8	1.25	.75	25-28	1.50	1.00
9-12	1.25	.75	29-32	2.50	1.00
13-16	1.25	.75	33-36	2.50	1.00
17-20	1.50	.75	37-40	2.50	1.00

Following are parts of the directions, and sample mazes are given in figures 9.1 and 9.2.

This is a test of your ability to plan a route between two points.

Look at the diagram below. (See figure 9.1.) Notice that there are four numbers, one at each corner of the maze. Each of these numbers is a starting point. The darkened square near the center of the maze is the common goal. Each number is connected with the goal by one or more lines. These lines are the routes that you must follow in going from each starting point to the goal.

Now study the various routes in the maze. Find the point or points through which you must pass in going from each starting point to the goal.

(After 20 seconds.) Turn the page. Now look at the maze below. (See figure 9.2.)

This maze is identical to the maze you have just studied except that there are letters on the various routes. Your task is to find the one letter through which you must pass in going from each starting point to the goal. In going from 91 to the goal, for example, you may pass through either A or C; however, you must pass through D; therefore D is the right answer. Now examine the route between 92 and the goal. Any route you follow takes you through B; therefore B is the right answer for 92. Similarly, in going from 93 to the goal, you must pass through G; therefore G is the right answer. In going from 94 to the goal, you may pass through either I or F; however, you must pass through D; therefore D is the right answer. (Note that D is the right answer for item 94 as well as for item 91.)

The test will proceed as follows. First, you will be shown a maze and told how long you will have to study it. After this study period, you will be told to turn the page and you will see a second maze, identical to the one you studied.

(3) *Scoring.*—The scoring formula is R−W/2.

Statistical results.—The data given below are for examinees of Psychological Research Unit No. 3.

Distribution statistics.—Available distribution data are given in table 9.2.

TABLE 9.2.—*Distribution constants for Route Planning, CI411AX*

Group	N	M	SD
Unclassified aviation students[1]	167	25.2	6.5
Classified pilots[2]	764	23.9	7.5

[1] Tested in May 1943.
[2] In Classes 44E, 44F, 44G, and 44H.

(2) *Reliability coefficient.*—An alternate-forms (part I-part II) reliability coefficient of 0.77, corrected, was obtained from a sample of 167 unclassified aviation students tested in May 1943. Since the two parts are not entirely comparable, this is a rough estimate.

(3) *Factorial composition.*—The chief loadings are in the planning (0.47), integration III (0.37), visualization (0.29), and general-reasoning (0.22) factors. The communality equals 0.63, which is somewhat short of its reliability (0.77).

(4) *Test validity.*—Validity for pilot training is indicated in table 9.3.

TABLE 9.3.—*Validity data for Route Planning, CI411AX, graduation-elimination criterion*

Group	N_t	p_s	M_s	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
Pilots in primary training[2]	764	0.88	24.01	23.02	7.47	0.07	0.15
Pilots through basic training[3]	455	.85	24.92	22.97	6.95	.15	.21

[1] Assuming an unrestricted stanine distribution of 2.00.
[2] In classes 44E, 44F, 44G, and 44H.
[3] In class 44F.

Evaluation.—Route Planning, CI411AX, has a validity for pilot success of approximately 0.16 that is fully accounted for by known factors as shown in chapter 28 (Table 28.18). It has no unique variance to offer except in the planning factor which has low validity for pilots. It is factorially complex and its known factors are better measured by other tests. The directions for route planning are relatively complex. It is therefore not a strong candidate for the air-crew classification battery.

Planning a Circuit, CI401A[6]

Planning a circuit presents a problem situation in which there is one and only one appropriate path to the objective, but a pathway that is obscured by a distracting maze of other pathways. An early form was developed under the title of Electrical Maze Test, CP401A.

Description.—Each item consists of an electrical-circuit diagram with many intersecting and intermeshed wires with several sets of terminals. The task is to trace the circuits and to determine at which pair of terminals a battery should be placed in order to complete the circuit through a meter.

(1) *Internal characteristics.*—The test contains 1 unrecorded and unscored sample problem, 2 recorded but unscored sample problems, and 42 scored items.

(2) *Administration.*—Fourteen minutes are allowed for completion of the test. After 12 minutes have elapsed, the administrator warns that only 2 minutes remain. Following are the directions and sample items. Fgure 9.3 is the sample problem utilized in the directions. Figure 9.4 is an example from the test proper, illustrating one of the more complex items.

Suppose that each of the following diagrams illustrates the wiring of the dashboard on an airplane. The small box at the top represents one of the meters on the panel. In order for the meter to work, a battery must be placed in the circuit at either A, B, C, D, or E. Only one of these points will successfully complete the circuit with but one battery. Your task is to find that place where a battery can be placed so that the meter will work; that is, which will complete a circuit through the meter.

From the example below, you can see that at one and only one place, such as C, can a battery be put in so as to complete the circuit successfully. All other choices, A, B, D, and E are incorrect; either they are connected with another point at which a battery would have to ' placed to make a complete circuit, or both wires from

[6] Developed at Headquarters, Army Air Forces. Chief contributors: Lt. Frank J. Dudek, Col. John C. Flanagan.

one of the points, A for example, go to the same pole of the meter. The effect of this is to short out a battery at that place so that it will not work. Dots in the figure represent connections; that is, the two wires are joined at that end. Where there is no such dot, the insulated wires simply cross each other but *no* connection is made.

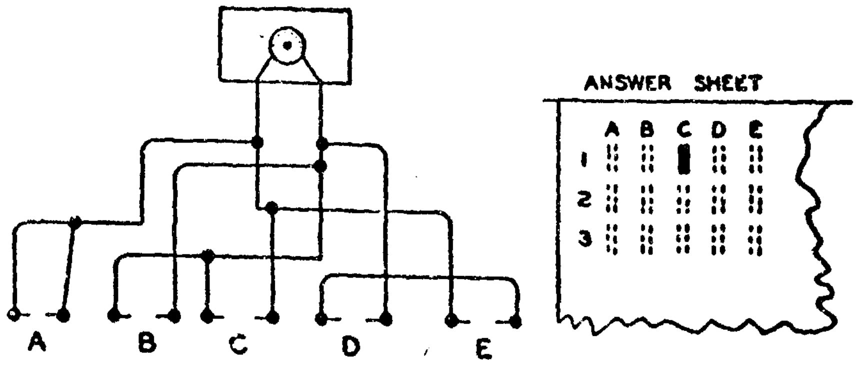


FIGURE 9.3
SAMPLE ITEM OF PLANNING A CIRCUIT, CI401A

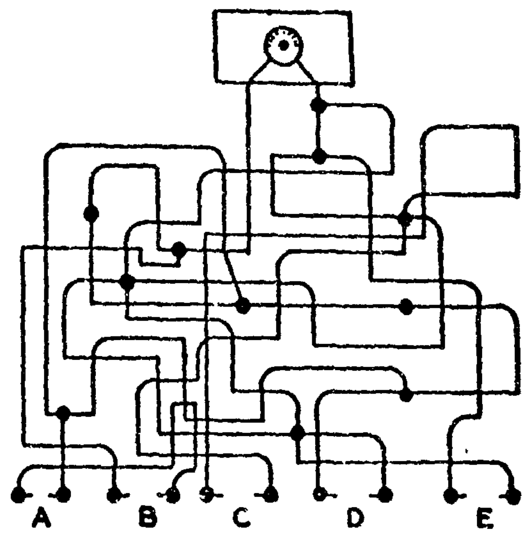


FIGURE 9.4
A DIFFICULT ITEM FROM PLANNING A CIRCUIT, CI401A

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results. (1) *Distribution statistics.*—Table 9.4 presents distribution data for two samples.

TABLE 9.4.—*Distribution statistics for Planning a Circuit, CI401A*

Group	N	M	SD
Unclassified aviation students[1]	197	23.3	8.5
Mixed[2]	669	30.1	7.9

[1] Tested in December 1942 at Psychological Research Unit No. 3.
[2] Sample of 508 unclassified aviation students, 156 pilot eliminees, and 5 unclassified students eliminated for medical reasons. Test administered with a 20-minute time limit in April 1942 at Psychological Research Unit No. 1.

(2) *Reliability coefficient.*—On the mixed sample of 669 cases (see footnote 2 to table 9.4), an estimated reliability coefficient of 0.96, corrected, was obtained by the odd-even method. Since this is a speed test, this figure is a serious overestimation.

(3) *Factorial composition.*—The leading factors and loadings are perceptual speed (0.41), planning (0.40), spatial relations (0.28), and verbal (0.24). The communality is 0.57. For a fuller picture of the factorial composition of this test, see appendix B.

(4) *Test validity.*—Validity data for pilots only are presented in table 9.5.

TABLE 9.5.—*Validity data for Planning a Circuit Test, CI401A and comparable forms, based upon graduation-elimination of pilots from primary training*

Test	N_s	p_g	M_g	M_e	SD_s	r_{bis}	$_cr_{bis}$[1]
CI401A	[2]222	0.79	24.95	22.34	7.86	0.19	...
QP901A[3]	598	.69	30.71	26.03	10.15	.28	.35
AC12I[3]	1,647	.64	27.11	25.37	6.77	.16	.22

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Tested in November and December 1942 at Psychological Research Unit No. 3.
[3] These contain the same 45 items. See report No. 6 on the AAF Qualifying Examination. Groups and testing dates are not identified.

Evaluation.—Planning a Circuit demonstrated relatively high validity for pilot success (composite mean of 0.26, based upon form CI401A and comparable forms), which is exactly accounted for by its known factors and their loadings (see table 28.18). Its factorial complexity, however, makes it undesirable except where its particular combination of factors is desired. This combination seems to coincide well with pilot requirements. It might, therefore, be used in a preliminary selective battery for pilots, such as the AAF Qualifying Examination.

TESTS OF ECONOMICAL PROCEDURES

In training and in combat, complex situations continually occur in which various alternative actions are possible. Several of the alternative actions may well lead to success. Success, in the sense of reaching the goal, however, is not sufficient. For although the goal may be achieved, the act of achieving may be too costly in terms of effort, time, or material. The pilot, bombardier, or navigator must engage in processes of selection—not only to select correct actions, but to select and execute the action which is most appropriate and most economical. He must foresee the shortest route, the fastest method, the simplest procedure. He must save time, effort, material.

A group of tests was designed to measure this ability to follow the most economical procedure in situations where various alternative actions are presented. These tests are Map Planning, Organizational Planning, Planning Air Maneuvers, and Sequence of Maneuvers.

Map Planning, CI412AX[7]

This is the first and only form of test by this name. It is the second

[7] Developed at Psychological Research Unit No. 3. Chief contributor: S/Sgt. Wayne S. Zimmerman.

of two tests designed to parallel the function of the Planning Maze Test, CI405A, a psychomotor test.

Description.—The examinee sees diagrammatic sections from city maps showing damage to streets following a bombing raid. The streets are blocked at various points by barriers represented as bomb craters. The examinee must plan routes for military vehicles to travel through the damaged areas. The task is to find the shortest passable route as quickly as possible.

(1) *Internal characteristics.*—Map Planning, CI412AX, contains four recorded but unscored sample items, all appearing in one diamond-maze sample map. There are 46 scored items in 5 mazes; 6 in the first maze, 8 in the second, 12 in the third, 10 in the fourth, and 10 in the fifth.

(2) *Administration.*—Each map or maze is timed separately with from 1.5 to 3.0 minutes per map being allowed. Total testing time including directions is 13 minutes.

Following are the directions and the sample items. The sample map (fig. 9.5) included is much reduced in size compared with the mazes found in the test proper.

This is a test of your ability to plan a route between two points. You will be shown sections from city maps showing damage to streets following a bombing raid. Assume that you must plan routes for military vehicles to travel through the damaged area. Your task will be to find the shortest passable routes as quickly as possible.

Look at the sample map below. Circles show places where falling bombs have rendered streets impassable. Note the numbers that appear on the margin of the map. Beginning with 1 at the upper left, the numbers go in a clockwise direction around the edges. These numbers indicate the points between which you must plan routes. Note, now, the small, square buildings within the map identified by letters of the alphabet. The shortest route between any two points will take you past one,

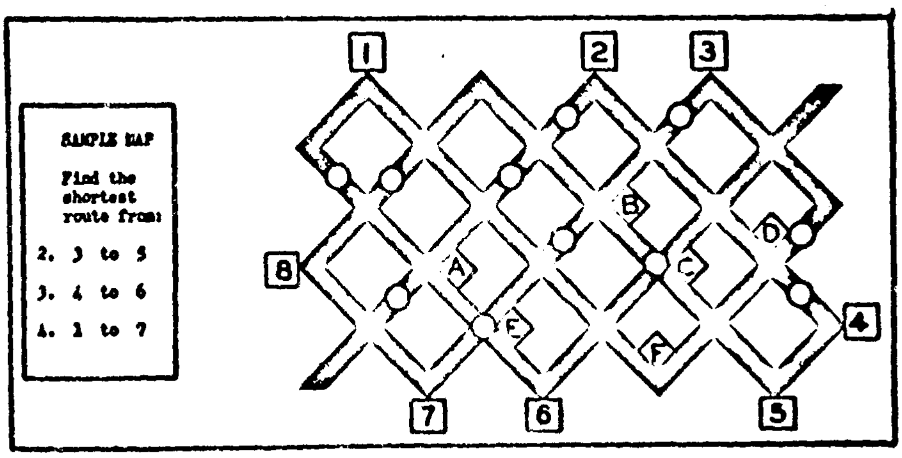


FIGURE 9.5
SAMPLE MAP OF MAP PLANNING, CI412AX

703320—47—12

and *only* one, of these lettered buildings. This is illustrated on the map by practice problem number 1 below:

Find the shortest route from:

1. 1 to 2.

Do that now.

The shortest route takes you by building B, so mark B on your answer sheet after item number 1. If you passed more than one building on your way, you did not find the shortest route. In every problem there is just *one* building on the shortest route between two numbered points. Work practice problems 2, 3, and 4 below. For items 2, 3, and 4 on your answer sheet, mark the letter corresponding to the building that you *must* pass.

(3) *Scoring.*—The scoring formula is R−W/2.

Statistical results.—The data for this test are limited but sufficient to permit an evaluation of its usefulness. The samples were tested at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Distribution constants are given in table 9.6.

TABLE 9.6.—*Distribution constants for Map Planning, CI412AX*

Group	N	M	SD
Unclassified aviation students[1]	167	26.4	6.9
Classified pilots[2]	864	20.2	6.9

[1] Tested in May 1943.
[2] In classes 44F and 44G.

(2) *Reliability coefficient.*—Correlating the scores on maps 1, 2 and 4, with the scores on maps 3 and 5, a reliability coefficient of 0.78, corrected, was obtained on a sample of 167 unclassified aviation students tested in May 1943.

(3) *Difficulty.*—The difficulty level of items in the test is indicated by the mean proportion of correct responses equal to 0.87, based on a group of 684 classified pilots. Standard deviation of the difficulty values was 0.09 and the range 0.53 to 0.97.

(4) *Factorial composition.*—The prominent loadings are in the perceptual-speed (0.45), general-reasoning (0.31), visualization (0.28), and spatial-relations (0.27) factors. The communality of 0.57 is sufficiently short of the reliability (0.78) as to indicate unknown common factors.

(5) *Test validity.*—Validation results based on several samples are given in table 9.7.

TABLE 9.7.—*Validity data for Map Planning CI412AX based upon samples of pilots, with graduation-elimination criterion*

Group	Class	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
In primary training	44F	404	0.91	20.42	18.20	7.17	0.16	0.25
Through basic training	44F	412	.89	20.17	18.73	7.22	.10	.17
In primary training	44G	460	.89	20.60	17.42	7.11	.23	.30
In primary training	44H	193	.85	25.47	25.65	6.51	−.02	.04
In primary training	44I	254	.82	18.29	18.08	6.60	.02	.07

[1] Assuming an unrestricted stanine standard deviation of 2.00.

(6) *Item validity.*—The validity of responses in this test is indicated by a mean phi of 0.08 with a range of phis from −0.11 to 0.28 and a standard deviation of 0.07. The data are based upon responses of 600 graduates and 84 eliminees in classes 44F and 44G.

Evaluation.—This test contains a number of valid factors which exactly account for its average pilot validity of 0.21. It has no unique variance to offer for pilot selection, but if it were not for its general reasoning component, it might still be used in a pilot-selection battery where pure tests are not demanded. The combination of factors is even better for navigator selection, for which it would probably validate to the extent of 0.30.

Organizational Planning, CI407BX [a]

This is the second and final form of another test in the economical-procedures subarea.

Description.—A schematic map of a town with various numbered buildings (post office, gas station, hardware store, etc.) is presented. The task is to plan and organize the shortest possible route which will include a series of stops. The examinee must foresee certain problems in connection with the most available and shortest routes and must plan accordingly.

(1) *Internal characteristics.*—The test contains one unrecorded and unscored sample problem and 42 scored items based on the map of a town.

(2) *Administration.*—Five minutes are required for the directions, and the time limit for the test is 50 minutes.

Following are the directions and sample problem. The map is shown in figure 9.6.

This is a test to see how well you can interpret a map. In some of the questions you will be asked to organize a trip to a series of places. You will have to foresee certain problems and plan accordingly, selecting the shortest or quickest route.

To help you locate points referred to in the questions, each place is given a number. Examine the map and note that these numbers are arranged in such a way that they get larger from left to right and from top to bottom. Note the ferry (34) at the lower right of the map. The ferry is toll free and runs every few minutes except where otherwise indicated.

Now work this sample problem:

You are at the bank (9). You want to stop at the following places before meeting a friend at the school (38). Which is the *first* place you will stop?

A. Shoe shop (11).
B. Docks (33).
C. Post office (6).
D. Bike shop (17).
E. Yacht club (23).

For this problem, *C* is the correct answer. The first stop on the shortest route is the post office (6).

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter, Jr., Lt. David H. Jenkins, Sgt. Betty J. Salk, S/Sgt. Wayne S. Zimmerman.

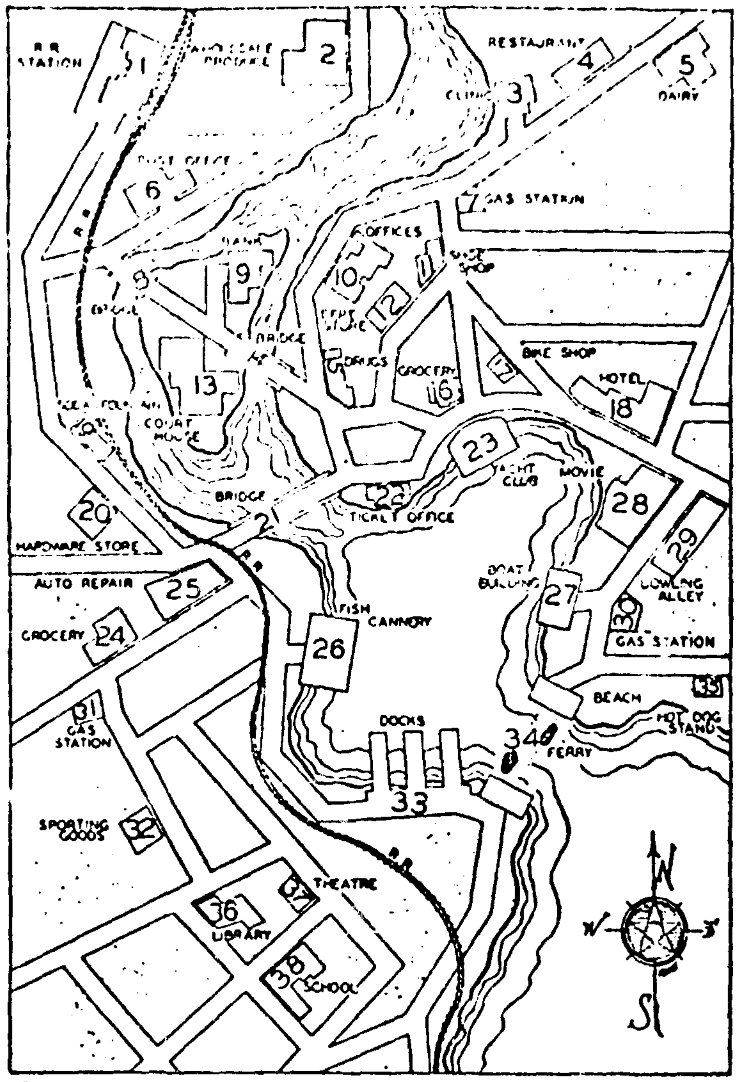


FIGURE 9.6
THE MAP USED IN ORGANIZATIONAL PLANNING, CI407BX

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—Data are fairly complete except for validity figures. The samples were tested at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Means and standard deviations are given in table 9.8.

TABLE 9.8.—*Distribution constants for Organizational Planning, CI407AX*[1] *and CI407BX, based on samples of unclassified aviation students*

Form	N	M	SD
CI407AX	[2]200	8.1	3.9
CI407BX	[3]275	16.7	6.1

[1] See page — for a de- tion of this variation.
[2] Tested in December 1942.
[3] Tested in May 1942 and August 1943.

(2) *Internal consistency.*—The internal consistency of items in this test is indicated by a mean phi of 0.33 with a range from 0.06 to 0.54 and a standard deviation of 0.14, based on the highest 25 percent and the lowest 25 percent of a sample of 200 unclassified aviation students tested with the CI407AX form in December 1942.

(3) *Reliability coefficients.*—These may be seen in table 9.9. They are somewhat low, though form B is apparently an improvement over form A.

TABLE 9.9.—*Reliability coefficients for Organizational Planning, CI407AX and CI407BX based on samples of unclassified aviation students*

Form	Method	N	r'_{11}	r_{11}
CI407AX	Part I vs. Part II	[1]200	0.29	0.45
CI407BX	Odd-even	[2]275	.50	.67

[1] Tested in December 1942.
[2] Tested in May 1942 and August 1943.

(4) *Difficulty.*—The difficulty level of items in form CI407AX is indicated by the mean proportion of correct responses equal to 0.34, with a standard deviation of 0.23 and a range of 0.00 to 0.88, based on the above-mentioned sample of 200 unclassified aviation students.

(5) *Factorial composition.*—The leading factor loadings of form CI407BX are in the numerical (0.38), integration II (0.35), integration III (0.28), and mechanical experience (0.20) factors. The communality (0.46) is sufficiently short of the estimated reliability (0.67) for this form to suggest room for other common factors. For a fuller picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Using a sample of 102 pilots in class 43I tested on the CI407AX form, a biserial correlation of 0.25, uncorrected, was obtained against the criterion of graduation-elimination from primary pilot training. The proportion of graduates was 0.76, the mean score of the graduates 8.82, the mean score of the eliminees 7.12, and the standard deviation of the scores of all was 3.96.

Evaluation.—Organizational Planning, CI407AX, has an uncorrected validity biserial of 0.25, based on a small sample. This validity estimate may be too high, since the N is small and the predicted validity based upon what is known of the test's factorial composition is only 0.18 (see table 28.18), but on the other hand the discrepancy may suggest unknown valid variance. Variance in the so-called planning factor is conspicuously absent, indicating that the test is misnamed. Its known factors are better measured by other much more reliable tests.

Variations of the test.—Organizational Planning, CI407AX, contained one unrecorded and unscored sample problem and 28 scored items divided into 2 parts of 14 each. All items were based on a map similar to the one used in the BX form. The time limit for part I was 22 minutes and for part II, 18 minutes. The high pilot validity biserial for CI407AX

(0.25) prompted the development of CI407BX. The number of items was increased and the map revised to appear more realistic.

Planning Air Maneuvers, CI408AX3 [a]

This is the third form of the third test in the economical procedures subarea. It was designed to measure the ability to visualize a course of action and to plan for its successful completion. The maneuvers in the test, as is often true of maneuvers in training or combat, must be made over the shortest, simplest, and most direct path.

Description.—This test assumes that the examinee is a sky-writing pilot who must plan how to write two adjacent letters by flying the shortest possible path. The starting and finishing positions of the plane are shown, and the sharpest turn that the plane can make is indicated. With this information and the large letters to be written presented in his test booklet, the examinee must select the correct path and indicate the direction in which he is traveling at each indicated point.

(1) *Internal characteristics.*—The test contains 18 recorded but unscored sample items and 87 scored items.

(2) *Administration.*—Twenty minutes are allowed for completing the test. After 10 minutes have elapsed, the examinees are informed that 10 minutes remain.

Following are part of the directions; the sample problem used in the directions is shown in figures 9.7 and 9.8.

This is a test of your ability to plan air maneuvers.

Assume that you are a sky-writing pilot and must plan how to write letter com-

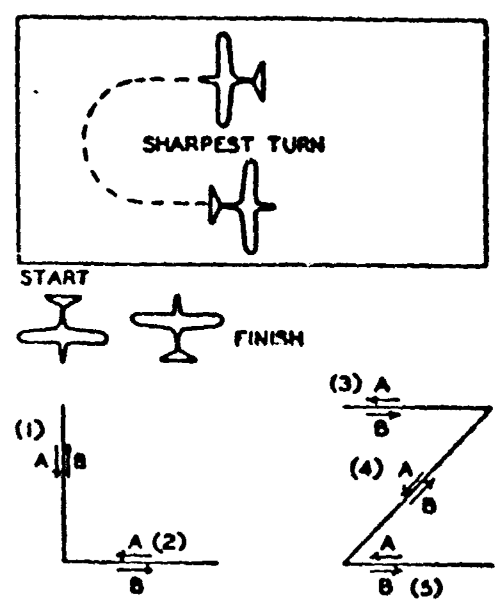


FIGURE 9.7
SAMPLE PROBLEM OF PLANNING AIR MANEUVERS, CI408AX3

[a] Developed at Psychological Research Unit No. 3. Chief contributor: S/Sgt. Wayne S. Zimmerman.

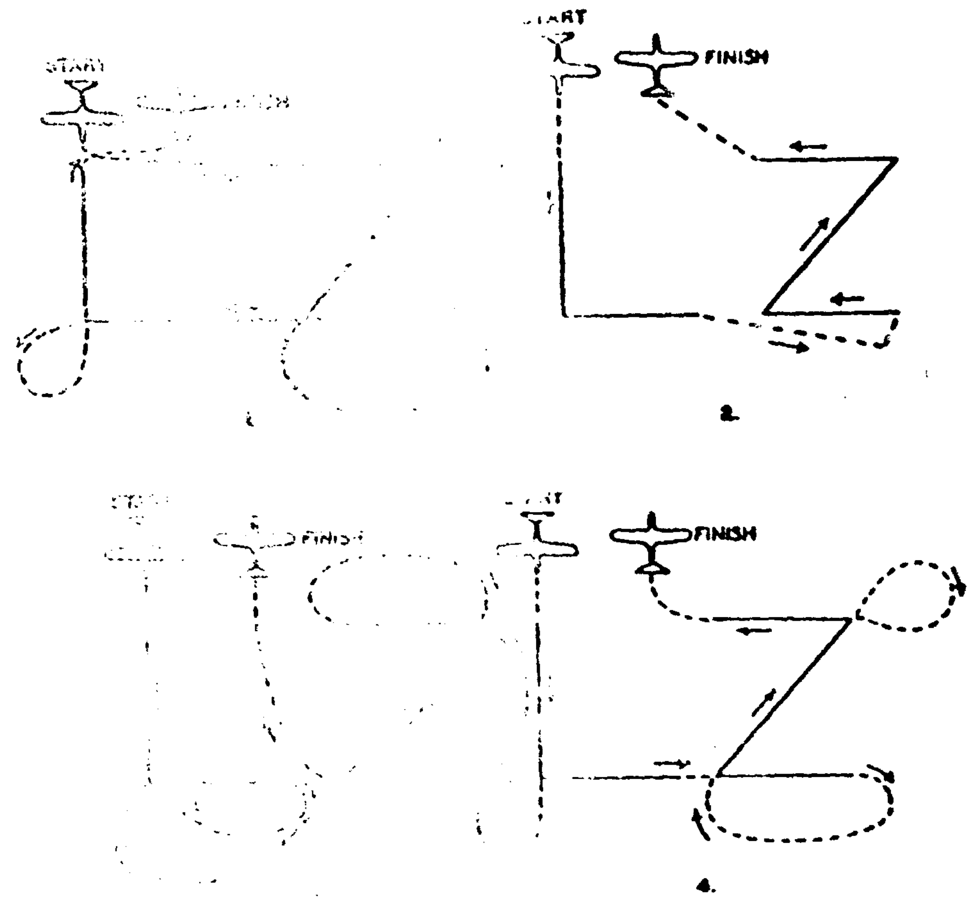
[illegible] maneuvers in the [illegible] problems that follow, you are to find the shortest, [illegible] order to do each problem correctly you must:

1. Be [illegible]
2. [illegible] second.
3. [illegible] plane is in the position labeled [illegible]
4. [illegible] more sharply than is shown in the [illegible] diagram and note the sharpest turn. (See [illegible])

Look at the [illegible] (fig. 9.7). In moving through these letters [illegible] You will indicate which direction you are [illegible] marking either A or B on your answer sheet.

Four [illegible] are illustrated. (See fig. 9.8.) Only one of [illegible] all of the rules set forth above. Decide [illegible]

Number [illegible] shortest, simplest, and most direct route that [illegible] first page. Illustration number 1 is wrong [illegible] before the first. Number 2 is wrong because [illegible] wrong because there is a shorter, simpler, and [illegible] maneuver.



[illegible] SOLUTIONS TO SAMPLE [illegible] MANEUVERS, [illegible]

Scoring.—The scoring formula is R — W.

Variations.—Since statistical data will be presented for three forms of the test, it is desirable to describe the variations here.

Planning Air Maneuvers, CI408AX1, contains 3 recorded but unscored sample items and 74 scored items divided into 2 parts of 35 each. Fifteen minutes are allowed for the completion of each part. The items in the AX1 form range greatly in difficulty because the combinations varied from one to four letters. As indicated by the statistical results, this form is too difficult and the items are not highly reliable. An effort to correct this situation was made in the revision, CI408AX2. This second form contains 5 recorded but unscored sample items and 114 scored items divided into 2 parts. Twenty minutes are allowed for the completion of each part. All sky-writing patterns in the AX2 form contain two letters, in contrast to the varying number (1–4) in AX1. Furthermore, only the letters easiest to trace, such as D, F, K, P, V, N, A, R, Z, L, H, and E are used. Other letters from the AX1 form that proved more difficult were dropped. Statistical data indicate that the revision achieved the desired effect.

In constructing form CI408AX3, items with the highest internal-consistency phis were taken from AX2 and the directions clarified. In form CI408BX1, the same items from AX3 are used. The only difference is that the directions were deliberately made brief. The purpose of the revision was to determine the effect of completeness and length of directions upon the functions measured by the test. No data are available.

Statistical results.—Data are available for the three forms of Planning Air Maneuvers, for examinees at Psychological Research Unit No. 3.

TABLE 9.10.—*Distribution data for AX1, AX2, and AX3 forms of Planning Air Maneuvers Test, for groups of classified pilots*

Test form	Number of items	N	M	SD
AX1	74	[1]227	20.3	11.9
AX2	114	[2]147	39.7	23.2
AX3	87	[3]1,148	31.6	16.[illegible]

[1] Tested in December 1942. Class not identified.
[2] In class 43K.
[3] In class 44F.

TABLE 9.11 *Internal-consistency data for Planning Air Maneuvers Test based on groups of unclassified aviation students*

Test form	N	Mϕ	SDϕ	Range of ϕ's
AX1	[1]675	0.25	0.14	−0.16 to 0.47
AX2	[2]675	.30	.13	−.11 to .64
AX3	[3]100	.44	.12	.18 to [4].65

[1] Tested in December 1942.
[2] Tested in February and March 1943.
[3] Tested in April and May 1943.
[4] Based on the total group for each item.

(1) *Distribution statistics.*—See table 9.10.

(2) *Internal consistency.*—Consistent improvement in test homo[illegible]ity is shown in th[illegible] three [illegible]ive forms. The data are presented [illegible] table 9.11.

(3) *Reliability coefficient.*—A reliability coefficient of 0.73, corrected, [illegible] obtained from a s[illegible] of 161 unclassified aviation students tested [illegible] March 1943. The coefficient was computed for Part I *v.* Part II of [illegible] AX2 form.

(4) *Difficulty.*—Th[illegible] [illegible] level of items in this test is indicated [illegible] the data in table 9.12.

[illegible] 9.12.—*Difficulty values corrected for chance success for Planning Air Maneuvers [illegible] on unclassified aviation students*

Form	N	M_p	SD_p	Range of p	
				Low	High
[illegible]X1	[1]675	0.47	0.23	0.00	0.90
[illegible]2	[2]675	.40	.21	.00	.76
[illegible]X3	[3]100	.45	.23	.00	.88

[1] T[illegible]ted in December 1942.
[2] Tested in February and M[illegible] 1943.
[3] T[illegible]ted in April and May 1943.

(5) *Factorial composition.*—Th[illegible] leading factor loadings for the AX3 [illegible] are in the planning (0.46), integration III (0.43), spatial-rela[illegible] (0.32), and mechanical experience (0.20) factors. The common[illegible] [illegible] variance is probably exhausted, with a communality of 0.69.

(6) *Test validity.*—Validity data are available for all three forms a[illegible] are presented in table 9.13.

T[illegible] 9.13.—*Validity data for three forms of Planning Air Maneuvers, CI408A, with the [illegible]ation-elimination criterion*

Group	Class	Test form	N_t	p_e	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
[illegible] in primary [illegible]	[2]...	AX1	227	0.81	[illegible]	25.05	11.94	0.28	...
[illegible] primary [illegible]	43K	AX2	147	.87	41.05	30.20	23.17	.25	...
P[illegible] in [illegible]imary [illegible]	44A	AX3	665	.82	26.30	23.20	15.30	.11	...
[illegible] in primary [illegible]	44F	AX3	1,1[illegible]	.94	32.55	29.42	16.44	.09	.15
[illegible]ough basic	[illegible]	[illegible]	1,131	.90	38.87	16.46	16.46	.14	.20

[1] [illegible] an unrestricted [illegible] standard deviation of 2.00.
[2] T[illegible] in December 1942, [illegible] not [illegible]ified.

Evaluation.—Plan[illegible] Air Maneuvers, CI408AX3, is strongly loaded with a new factor, which has been difficult to define. In three different analyses, its loading [illegible] been 0.51, 0.46, and 0.33, with a mean of 0.46. [illegible] this factor [illegible] in conjunction with Planning a Circuit and [illegible] judgment tests, it is probable that this factor has a small positive [illegible]ity for pilots and that it contributes to the average pilot validity [illegible] obtained for the form of the Planning Air Maneuvers test. Its [illegible] in integrati[illegible] III is probably a handicap in relation to pilot validity. Its pilot validity is fully accounted for by known factor composition, assuming a validity f −0.25 for integration III, which detracts 0.10 from the total estimate. This calls for steps to rid the test of its integration III variance.

Sequence of Maneuvers, CI410A [10]

This is another test in the economical procedures subarea. It attempts to duplicate on paper a type of planning that many pilots must execute in the course of their training and operations.

Description.—The examinee is presented with diagrams of a series of five maneuvers involving climbs and dives. He must take into account the altitude at which each maneuver must be done and the amount of altitude lost or gained. Then he must plan the arrangement for executing the five maneuvers so as to do the least amount of unnecessary climbing and diving.

(1) *Internal characteristics.*—The test contains 2 recorded but unscored sample items and 32 scored items, divided into 2 parts of 16 each.

(2) *Administration.*—Twenty-seven minutes are allowed for the completion of each part. Six minutes are required for directions, bringing the total testing time to 60 minutes. Following are the directions and sample items. Purely oral directions are given entirely in italics.

Suppose you have had several hours of solo flying in primary flying school and your instructor tells you to take up a plane and do several maneuvers. You would have to figure out in what order you would do the maneuvers, for surely you would not attempt an [illegible]ide loop just after having completed a power dive which left you quite near the [illegible]nd, would you? This is a test to measure just such an ability, to see how well you can plan air maneuvers in flying.

Now look at the cover sheet of your booklet and read the directions silently as I read them aloud.

This is a test of your ability to plan the most efficient order in which to carry out a series of practice air maneuvers. For each maneuver in a series, you will be told the altitude lost or gained while performing it. You will also be told when a given maneuver should be carried out between two definite altitudes, or when it must begin or end above or below a particular altitude. For each series of five maneuvers, you will be told the altitude at which you are to start the series. Your problem is to figure out the most efficient order in which to perform each set of maneuvers, i. e., the order that involves the least amount of unnecessary climbing and diving. You are to note which of the maneuvers comes ***fourth*** in your arrangement, and indicate your answer by marking in the space on your answer sheet under the letter which corresponds to the maneuver that you decide should come ***fourth*** in the sequence. ***You may climb or dive before the first maneuver and between maneuvers, but you should change altitude as little as possible before and between maneuvers.***

Look at example 1. See figure 9.9.

Note that in maneuver A the plane loses 3,000 feet in altitude and must finish the maneuver at above 2,000 feet. The maneuver can end at any altitude above 2,000 feet so long as it is begun at an altitude 3,000 feet higher. Similarly, ma-

[10] Developed at Psychological Research Unit No. 3. Chief contributor: Lt. Mahlon B. Smith.

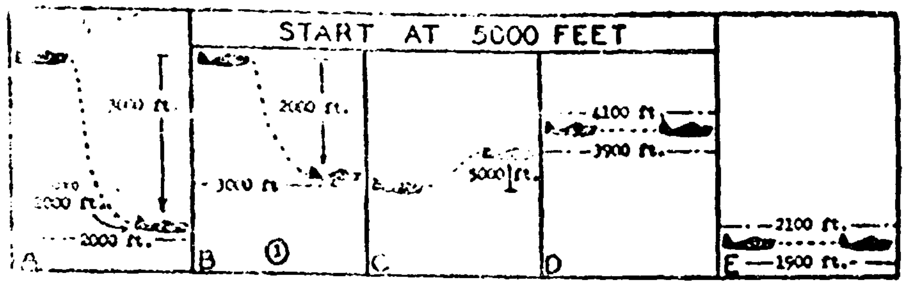


FIGURE 9.9
SAMPLE PROBLEM OF SEQUENCE OF MANEUVERS, C1410A

...uver B involves a loss of 2,000 feet and must be completed at an altitude of 3,000 ... or above. Maneuver C requires a 500 foot climb, but *may be performed at any altitude.* In maneuvers D and E, there is no change in altitude but the entire maneuver must be done between the altitude levels indicated.

Now look at me. (Administrator should pause until all heads are lifted.) ***At this point I will give you some additional explanation that is not printed in your test booklet. Look at maneuver D in example 1. The lines at 3,900 and 4,100 feet mean that this maneuver must be done at approximately 4,000 feet. Similarly, the lines ... E mean that maneuver E must be done at approximately 2,000 feet. Just to the right of B you will find the number 1 in a small circle.*** (Administrator should pause briefly to allow cadets to find the encircled number.) ***This means that maneuver B is the first maneuver to be done when all the maneuvers are arranged in the proper order.***

Now we will work out the first example together. Notice that at the top of example one it states that we must start at 5,000 feet. We already know that the first maneuver should be B. This brings us down from 5,000 feet to 3,000. Which one should we do next? Look over the four maneuvers that are left. Notice that A requires the highest altitude, for it also must be begun at an altitude of at least 5,000 feet. Maneuver A involves the greatest loss of altitude of any of the maneuvers, and will bring us in position to perform maneuver E, at 2,000 feet. We must therefore get from 3,000 feet, where maneuver B left us, to the 5,000 feet required to do maneuver A. This involves a climb of 2,000 feet, and we can use maneuver C for part of this, doing C on the way up. C is thus our second maneuver, and brings us to 3,500 feet. D must be done at an altitude of about 4,000 feet, so we can do that on our way up to 5,000 feet without losing any altitude. We will therefore climb 500 feet more, do maneuver D, and then climb another thousand feet to 5,000 feet. Now we are in position to do maneuver A, which brings us down from 5,000 feet to 2,000 feet. This is our fourth maneuver. At 2,000 feet we can perform maneuver E, our fifth maneuver, without any further change of altitude. We selected each maneuver with the aim of getting into the best possible position to perform the rest of the maneuver.

Now we will continue reading the directions in the test booklet.

The least amount of extra climbing and diving is therefore involved when the maneuvers are performed in the order B, C, D, A, E. A is the correct answer, i.e., the fourth maneuver in the proper sequence. The diagrams below show why this is the best arrangement. See figure 9.10.

Diagram I shows how the maneuvers may be performed most efficiently, as we have just done them, while diagram II shows one of the less efficient, incorrect solutions. The dotted lines represent the maneuvers, which are labeled with the same letters that they had in the first diagram. The solid lines represent changes

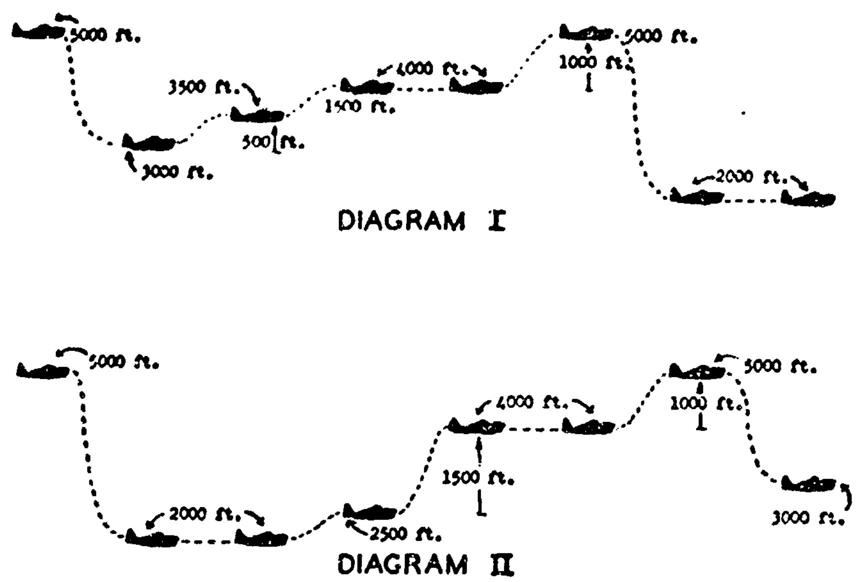


FIGURE 9.10
CORRECT AND INCORRECT SOLUTIONS TO SAMPLE PROBLEM OF SEQUENCE OF MANEUVERS, CI410A

in altitude necessary to get into position to perform maneuvers. When the maneuvers are done in the correct order shown in diagram I, a 500-foot climb between maneuvers C and D and a 1000-foot climb between maneuvers D and A are necessary. Any other order would require more climbing or diving. For example, when the maneuvers are done in the order A, E, C, D, B, as in diagram II, a 1,500-foot climb between maneuvers C and D and a 1,000-foot climb between maneuvers D and B are necessary. *The second arrangement is poor because maneuver A, which involves the greatest loss of altitude, is performed first instead of B, leaving the plane in a poor position to perform maneuvers D and B, both of which require relatively high altitudes.* Maneuver A is fourth in the best arrangement, so blacken the space A after number 1 on your answer sheet.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—Data are limited for this test, but are sufficient to afford some evaluation of it. The samples upon which the data are based were tested at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—The distribution of scores in this test is indicated by a mean of 10.3 and a standard deviation of 5.6, based on a sample of 436 unclassified aviation students tested in December 1942.

(2) *Internal consistency.*—The internal consistency of items in this test is indicated by a mean phi of 0.39, with a range from 0.10 to 0.57 and a standard deviation of 0.11, based on the highest 27 percent and the lowest 27 percent of 220 unclassified aviation students.

(3) *Reliability coefficient.*—An alternate-form reliability coefficient of 0.66, corrected, was obtained from the above-mentioned sample of 436 unclassified aviation students.

(4) *Difficulty.*—The difficulty level of items in the test is indicated by a mean proportion of correct responses equal to 0.32, corrected for

chance success, with a standard deviation of 0.12 and a range from 0.02 to 0.62, based on a sample of 220 unclassified aviation students.

(5) *Factorial composition.*—The leading factor loadings are in the verbal (0.39), judgment (0.38), planning (0.35), and numerical (0.30) factors. Its communality (0.59) indicates that practically all of its non-error variance (0.66) is known.

(6) *Test validity.*—For a sample of 247 pilots in primary training, originally tested in December 1942, the validity coefficient was 0.00. The proportion of graduates was 0.77, the means of graduates and eliminees were 9.62 and 9.64, respectively, and the standard deviation of all was 5.30.

Evaluation. Sequence of Maneuvers, CI410A, is not considered suitable for administration because of its extremely complicated directions. Furthermore, the reliability coefficient is relatively low, 0.66, corrected, which is even more serious in view of the length of time required to administer the test in its present form. The obtained validity for pilots in a small sample was 0.00, and from its factor loadings one would not expect a pilot validity greater than 0.10. It combines two factors valid for navigators—verbal and numerical, but the validities of its other two factors for navigator selection are unknown.

A TEST OF PLANNING BY DEDUCTION

It is reasonable to suppose that the victor in aerial combat is usually the pilot who can anticipate his opponent's moves and then plan his own maneuvers accordingly. In such planning, the pilot is aware of certain general factors that govern his action; such factors, for example, as the position of clouds and the limitations of his own and of his opponent's airplane. Thus, from observation of the situation, the pilot must *plan by deduction* what his opponent will probably do, and then what he can do to gain advantage. The attempt was made to embody this deductive aspect of planning in a test called competitive planning.

Competitive Planning, CI409AX2 [11]

This is the final form of the only test in the subarea planning by deduction.

Description.—This test is based on the familiar Completion-of-Squares game, sometimes called "Squares" or "Boxes." In the test, the examinee must plan moves for both opponents, so that each completes as many squares as possible in a rectangular diagram of incomplete square figures. In order to solve the problems correctly, the examinee must anticipate the best moves for each opponent. The most attractive immediate move is not always the best move. It was felt that it would be desirable in at least one foresight-and-planning test to provide an opportunity for the examinee to refuse immediate gains in favor of later benefits.

[11] Developed at Psychological Research Unit No. 3. Chief contributors: S/Sgt. J. Gordon Ekin, and Lt. Lion Hutchinson.

(1) *Internal characteristics.*—The test contains 1 unrecorded and unscored sample problem, 2 recorded but unscored practice problems, and 40 scored items divided into two parts of 20 each. Each diagram is presented in duplicate so that the examinee may try a second solution without erasing his first attempts.

(2) *Administration.*—Solutions to problems are worked out by marking lines or completing squares directly on the work booklet. These solutions are then entered on the standard five-place IBM answer sheet. Seventeen minutes are allowed for the completion of each part.

Following are parts of the directions contained in the test booklet. The practice diagrams referred to appear in the work booklet along with the scored items of the test. They are shown in figure 9.11.

This is a test to see how well you can plan moves in a competitive situation.

Examine the diagrams on page 1 of the work book. Two contestants, Black and White, took turns filling in the sides of incompleted squares in patterns similar to those shown on page 1 of the work book. Each of the contestants always made the best possible moves for himself. Your task will be to reconstruct the moves made by the two contestants.

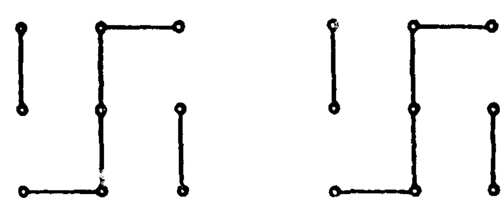
The rules were as follows—read them carefully:

a. Black always made the first move, filling in one side of an incompleted square.

ANSWER LEGEND

A -	BLACK 0	WHITE 4
B -	BLACK 1	WHITE 3
C -	BLACK 2	WHITE 2
D -	BLACK 3	WHITE 1
E -	BLACK 4	WHITE 0

SAMPLE PROBLEM



PRACTICE PROBLEMS

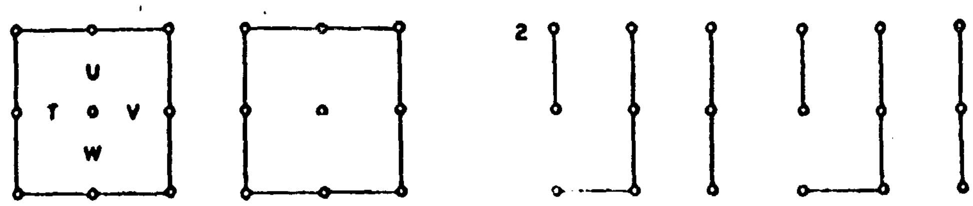


FIGURE 9.11
SAMPLE AND PRACTICE PROBLEMS OF COMPETITIVE PLANNING C1409AX2

b. Each time Black or White completed a square, he had to, make one additional move. A square is completed when the fourth side is filled in.

c. Each opponent completed the greatest possible number of squares in the finished pattern.

You are to work each problem making only those moves which, at the end of the problem, give each competitor the highest possible scores despite the best efforts of the other. Your answer is the number of squares each contestant completed in the finished pattern. If you fail to select the best possible moves for both of the opponents, you will not get the correct answers.

You may mark in the work book in order to solve the problem. Each diagram is given in duplicate so that you may try a second solution without erasing.

Now study the sample problem on page 1 of the work book.

Making the best moves possible, Black and White completed this sample problem in the following manner: As always, the first move was made by Black. No matter which side Black filled in, White was able to complete two squares immediately. After completing his second square, he, White, was compelled, by the rules, to fill in a side on one of the squares in the other half of the pattern. This enabled Black to complete the remaining two squares so that the final result became two squares for Black and two squares for White. Since this result is listed opposite choice C in the list of alternate answers at the top of the page, the correct answer to this problem would be C.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data available are for samples tested in April and May 1943, at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Means and standard deviations yielded by two samples are shown in table 9.14.

TABLE 9.14.—*Distribution constants for Competitive Planning, CI409AX2*

Group	N	M	SD
Unclassified aviation students	422	21.4	7.4
Classified pilots[1]	682	20.1	6.4

[1] In classes 44D and 44E.

(2) *Internal consistency.*—The internal consistency of items in this test is indicated by a mean phi of 0.38 with a range from −0.17 to 0.80 and a standard deviation of 0.21, based on the highest 27 percent and the lowest 27 percent of 422 unclassified students.

(3) *Reliability coefficient.*—A reliability coefficient of 0.68, corrected, was obtained by the part I-part II method on a sample of 422 unclassified students.

(4) *Difficulty.*—The difficulty level of items in the test is indicated by the mean proportion of correct responses equal to 0.51, corrected for chance success, with a standard deviation of 0.25 and a range from 0.00 to 0.96, based upon the above-mentioned sample of 422 cases.

(5) *Factorial composition.*—The chief loadings are in these factors: general-reasoning (0.36), judgment (0.36), and integration III (0.33), with a slight contribution from visualization (0.19). The communality (0.48) falls short of the reliability (0.68).

(6) *Test validity.*—Using a sample of 682 pilots in classes 44D and 44E, a biserial correlation of 0.19, corrected, was obtained with the criterion of graduation-elimination from primary training. The proportion of graduates was 0.92, the mean score of the graduates 20.26, the mean score of eliminees 18.22, and the standard deviation of all scores was 6.39.

Evaluation.—Competitive planning has a validity (0.19) for pilots that is largely unaccounted for by factors of known pilot validity. In fact, the discrepancy between predicted validity (0.05) and the obtained is so great as to justify search for the unknown valid components. The test is probably handicapped by its variance in integration III and should be freed from that element. The general-reasoning factor also contributes excess variance, which could well be dispensed with so far as pilot selection is concerned.

Variations—The CI409AX1 form of competitive planning contained only 20 items divided into two parts of ten items each. Directions and problems were consolidated into one booklet. Some items in this earlier form were considerably more difficult than those in the later revision. Whereas the X2 items never exceed four squares, the X1 items were graduated in difficulty from two to nine squares. It was believed that difficulty was entirely a function of the number of squares. Total testing time required for this preliminary form was 40 minutes (including directions). The X1 form had low reliability (part I *v.* part II reliability corrected was only 0.28 on a sample of 200 unclassified students tested in December 1942).

A FACTOR ANALYSIS OF FORESIGHT AND PLANNING TESTS [12]

Analyses were made of two special batteries of foresight and planning tests [13] in order to try to understand better their fundamental variances and to test the hypothesis that there are such fundamental human abilities as foresight and planning or a single factor underlying the two.

The Data

The two batteries include a small number of planning tests plus a few tests selected from the classification battery because of their recognized reference value, plus some experimental tests in the areas of reasoning and judgment. One basis for the inclusion of judgment tests was that in certain judgment items which contain problems of a work-planning sort there seemed to be a unique variable. All printed tests involved in the analyses are described in this chapter or elsewhere in this volume. The one psychomotor test—Complex Coordination—is described briefly on p. 122 and more completely in Report No. 4.

[12] Executed by S/Sgt. J. Gordon Etkin, S/Sgt. Benjamin Fruchter, Capt. Lloyd G. Humphreys, Lt. David H. Jenkins, Sgt. Harold H. Singer, and S/Sgt. Wayne S. Zimmerman at Psychological Research Unit No. 3.
[13] Hereafter called planning tests, for convenience.

Table 9.15.—*Correlation matrix for the foresight and planning battery I* ($N=202$)[1]

	Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Mechanical Judgment[2]	..	40	32	24	33	25	36	16	42	55	54	39	21	23	21	28
2	Non-mechanical Judgment	40	..	25	40	30	32	39	13	33	25	36	37	33	19	15	25
3	Organizational Planning	32	25	..	39	47	38	39	12	25	26	33	42	44	18	32	33
4	Sequence of Maneuvers	24	40	39	..	42	44	39	12	17	16	21	49	50	15	09	26
5	Spatial Reasoning	33	30	47	42	..	46	40	20	22	16	30	45	47	21	23	39
6	Planning Air Maneuvers	25	32	38	44	46	..	44	17	26	26	40	44	37	02	[illegible]	[illegible]
7	Planning a Circuit	36	39	39	39	40	44	..	31	24	34	39	33	22	32	[illegible]	40
8	Pursuit Test	16	13	12	12	20	17	31	..	18	15	08	03	08	31	37	36
9	Driving Skill	42	33	25	17	22	26	24	18	..	42	52	27	20	15	12	37
10	Mechanical Information	55	25	26	16	16	26	34	15	42	..	61	50	07	04	07	25
11	Mechanical Principles	54	36	33	21	30	40	39	08	52	61	..	53	28	07	[illegible]	[illegible]
12	Reading Comprehension	39	37	42	49	45	44	33	03	27	50	53	..	46	11	[illegible]	[illegible]
13	Mathematics B	21	33	44	50	47	37	22	[illegible]	20	07	28	46	..	08	[illegible]	[illegible]
14	Speed of Identification	23	19	18	15	21	02	32	31	15	04	07	11	08	..	47	22
15	Spatial Orientation I	21	15	32	09	23	14	39	37	12	07	06	09	18	47	..	23
16	SAM Complex Coordination	28	25	33	26	39	35	40	30	37	26	44	20	09	22	23	..

[1] Decimal points omitted.
[2] For code numbers see table 9.16.

Table 9.16.—*Centroid loadings and communalities for foresight and planning battery I*[1]

	Test	I	II	III	IV	V	VI	VII	VIII	IX	h^2
1	Mechanical Judgment, CI301BX1	62	17	29	03	19	−11	−10	15	03	58
2	Non-mechanical Judgment, CI301BX1	55	16	03	28	−16	−12	−12	−09	−05	47
3	Organizational Planning, CI407AX1	61	−06	−24	−09	13	−07	06	10	−18	51
4	Sequence of Maneuvers, CI410A	56	10	−38	18	−19	−03	−14	07	−21	61
5	Spatial Reasoning, CI211BX1	62	−10	−34	−08	03	−15	−05	−12	04	56
6	Planning Air Maneuvers, CI408AX1	59	10	−25	−17	−26	15	02	−15	−07	57
7	Planning a Circuit, CI401A	66	−19	05	03	−14	17	−16	−05	−06	56
8	Pursuit, CP512A	35	−39	16	04	−13	12	13	−10	−07	36
9	Driving Skill, CI307AX1	53	19	32	−09	−06	−22	21	−03	−04	53
10	Mechanical Information, CI905A	54	33	37	−19	08	27	−10	14	−10	69
11	Mechanical Principles, CI903A	65	36	27	−26	06	−03	−07	−18	05	74
12	Reading Comprehension, CI614B	64	37	−20	−02	26	14	−20	−10	−11	74
13	Mathematics B, CI206B	51	14	−47	16	14	−15	19	−12	−07	62
14	Speed of Identification, CP610A	37	−46	17	25	13	−10	−11	−07	06	49
15	Spatial Orientation I, CP501B	41	−53	06	17	20	12	19	−09	−08	59
16	SAM Complex Coordination, CM701A	55	−18	12	−31	−22	−19	−09	−07	−07	55

[1] Decimal points omitted.

703320—47—13

TABLE 9.17.—*Rotated factor loadings for foresight and planning battery I*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	h^2
1 Mechanical Judgment[2]	20	07	15	05	29	12	54	36	−01	59
2 Nonmechanical Judgment	16	15	13	01	30	17	13	39	36	49
3 Organizational Planning	19	41	29	29	02	27	20	13	05	50
4 Sequence of Maneuvers	00	30	21	19	−00	39	00	38	35	59
5 Spatial Reasoning	18	19	44	33	19	34	02	14	12	56
6 Planning Air Maneuvers	00	16	30	23	10	26	17	07	51	54
7 Planning A Circuit	41	02	10	28	08	24	24	19	40	57
8 Pursuit	52	19	09	02	−01	01	07	−01	27	39
9 Driving Skill	08	17	11	15	42	−05	46	21	15	53
10 Mechanical Information	03	00	−06	09	12	20	77	09	17	70
11 Mechanical Principles	00	01	19	19	50	20	61	05	16	76
12 Reading Comprehension	−02	20	23	01	20	65	39	13	17	75
13 Arithmetic Reasoning	03	48	48	−01	13	32	−04	10	13	61
14 Speed of Identification	65	01	02	01	14	05	−02	22	04	50
15 Spatial Orientation I	69	24	11	05	−05	05	03	−05	13	58
16 SAM Complex Coordination	19	04	11	56	28	02	20	11	18	53

[1] Decimal points omitted.
[2] For code numbers see table 9.16.

TABLE 9.18.—*Correlation matrix for Foresight and Planning Battery II (N=170)*[1]

	Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	Speed of Identification[2]	...	[3]47	[3]11	[3]12	[3]05	[3]07	[3]08	[3]14	[3]21	[3]10	17	19	26	[3]10	06	−02	01	11	21
2	Spatial Orientation I	[3]47	...	16	[3]13	[3]17	[3]25	[3]03	[3]12	[3]27	[3]15	30	15	36	[3]15	15	09	−03	12	19
3	Technical Vocabulary—Nav	[3]11	[3]16	...	[3]56	[3]38	[3]20	[3]15	[3]29	[3]12	[3]29	24	−03	01	[3]16	22	12	14	12	15
4	Reading Comprehension	[3]12	[3]13	[3]56	...	[3]43	[3]10	[3]42	[3]56	[3]22	[3]37	02	04	01	[3]20	40	32	22	24	20
5	Mathematics B	[3]05	[3]17	[3]38	[3]43	...	[3]52	[3]11	[3]29	[3]16	[3]32	43	17	27	[3]35	33	41	17	[illegible]	14
6	Numerical Operations	[3]07	[3]25	[3]20	[3]10	[3]52	...	[3]−14	[3]−06	[3]11	[3]16	23	05	20	[3]25	20	14	08	07	07
7	Mechanical Information	[3]08	[3]03	[3]15	[3]42	[3]11	[3]−14	....	[3]55	[3]22	[3]20	00	−07	−04	[3]−05	19	20	20	[illegible]	[illegible]
8	Mechanical Principles	[3]14	[3]12	[3]29	[3]50	[3]29	[3]−06	[3]55	....	[3]32	[3]36	19	21	12	[3]27	37	31	25	[illegible]	24
9	SAM Complex Coordination	[3]21	[3]27	[3]12	[3]22	[3]16	[3]11	[3]22	[3]32	...	[3]30	44	05	29	[3]12	27	19	11	[illegible]	11
10	Planning Air Maneuvers	[3]10	[3]15	[3]20	[3]37	[3]32	[3]16	[3]20	[3]36	[3]30	...	38	41	12	[3]23	28	16	25	[illegible]	24
11	Planning a Course	17	30	24	02	43	23	00	19	44	38	..	25	37	37	24	24	07	[illegible]	21
12	Route Planning	19	15	−03	04	17	05	−07	21	01	41	25	..	39	31	17	15	15	[illegible]	11
13	Map Planning	26	36	01	01	27	[illegible]	−04	12	29	12	37	39	...	27	25	20	19	14	26
14	Competitive Planning	[3]10	[3]15	[3]16	[3]20	[3]35	[3]25	[3]−05	[3]27	[3]12	[3]23	37	31	27	....	27	29	19	10	10
15	Practical Judgment I	06	15	32	40	31	20	19	37	27	28	28	17	25	27	..	30	35	32	31
16	Practical Judgment II	−02	09	12	32	41	14	20	31	19	16	24	15	20	29	30	..	32	18	09
17	Practical Estimations I	01	−03	14	22	17	08	20	25	11	25	03	15	10	19	35	32	....	26	16
18	Practical Estimations II	11	12	12	24	22	09	20	34	10	35	16	18	14	10	32	18	26	..	11
19	Judgment of Proportions	21	19	15	20	14	07	06	24	11	24	21	11	26	10	31	09	16	11	..

[1] Decimal points omitted.
[2] For code numbers see table 9.19.
[3] These coefficients were taken from more stable data based on the same type of examinees

TABLE 9.19.—*Centroid loadings and communalities for the Foresight and Planning Battery II*[1]

	Test	I	II	III	IV	V	VI	VII	VIII	IX	X	h^2
1	Speed of Identification, CP610A	33	34	42	05	15	09	15	09	11	07	48
2	Spatial Orientation I, CP501B	42	40	31	−13	18	−04	09	13	08	03	52
3	Technical Vocabulary (Nav), CE505D	45	−26	15	−36	08	17	−18	−20	03	11	54
4	Reading Comprehension, CI614G	58	−48	25	−17	−05	23	−08	−06	10	−16	75
5	Mathematics B, CI706, CI206B	62	−08	−30	−43	−06	−09	−06	05	29	08	77
6	Numerical Operations, CI701B	35	15	−26	−39	21	−04	−07	19	07	06	46
7	Mechanical Information, CI905A	33	−46	32	24	−19	−17	03	18	06	05	58
8	Mechanical Principles, CI903A	61	−35	22	24	−27	−04	09	−10	10	16	73
9	Complex Coordination, CM701A	46	13	15	08	−18	−24	−09	06	−17	−05	39
10	Planning Air Maneuvers, CI405AX3	57	04	−06	17	−24	29	−32	16	09	−18	67
11	Planning A Course, CI406AX2	54	37	−10	−10	−24	−20	−27	−09	−16	14	67
12	Route Planning, CI411AX1	38	29	−25	29	−12	28	22	−15	17	−12	58
13	Map Planning, CI412A``1	45	41	−08	11	08	−24	18	−16	13	−13	54
14	Competitive Planning, CI409AX2	46	13	−26	−12	−13	09	21	−23	−11	15	47
15	Judgment of Proportions, CP206B	37	07	07	14	20	−04	−10	−20	06	−13	28
16	Practical Judgment I, CI301BX3	59	−16	−09	08	19	−07	−10	−18	−12	−05	49
17	Practical Judgment II, CI301BX3	47	−21	−25	−09	−13	−18	26	04	−17	−14	50
18	Practical Estimations I, CI308AX1	38	−24	−21	24	16	07	09	10	−18	−09	39
19	Practical Estimations II, CI308AX1	42	−14	−12	24	11	10	−08	14	08	14	34

[1] Decimal points omitted.

TABLE 9.20.—*Rotated factor loadings for the Foresight and Planning Battery II*[1]

Test[2]	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
1 Speed of Identification	65	09	−08	[illegible]	[illegible]	12	[illegible]	−05	01	[illegible]	[illegible]
2 [illegible]	64	[illegible]	02	[illegible]	[illegible]	[illegible]	05	02	[illegible]	[illegible]	[illegible]
3 Technical Vocabulary (Navigation)	−02	28	[illegible]	[illegible]	[illegible]	[illegible]	14	[illegible]	−05	[illegible]	[illegible]
4 Reading Comprehension	01	[illegible]	[illegible]	[illegible]	02	69	49	04	18	[illegible]	[illegible]
5 Mathematics B	−03	65	50	[illegible]	14	27	[illegible]	04	06	07	[illegible]
6 Numerical Operations	[illegible]	63	[illegible]	09	−06	09	[illegible]	14	[illegible]	[illegible]	[illegible]
7 Mechanical Information	00	[illegible]	07	[illegible]	[illegible]	11	74	−10	[illegible]	−10	[illegible]
8 [illegible]	05	[illegible]	[illegible]	[illegible]	[illegible]	26	[illegible]	00	09	[illegible]	[illegible]
9 [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	07	[illegible]	10	[illegible]	[illegible]	[illegible]
10 [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11 Planning [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
12 Route Planning	24	[illegible]	[illegible]	[illegible]	[illegible]	−05	[illegible]	[illegible]	47	[illegible]	[illegible]
13 Map Planning	45	12	31	27	28	−05	−09	16	24	[illegible]	[illegible]
14 Competitive Planning	07	19	29	03	32	12	−02	36	01	33	[illegible]
15 Judgment of Proportions	21	04	00	20	22	22	02	[illegible]	30	−09	22
16 Practical Judgment I	02	20	03	24	29	30	21	[illegible]	28	−07	[illegible]
17 Practical Judgment II	−05	15	40	10	04	03	32	45	08	01	51
18 Practical Estimations I	−05	[illegible]	−04	00	07	01	33	36	36	01	[illegible]
19 Practical Estimations II	00	28	−08	02	22	03	32	02	31	11	[illegible]

[1] Decimal points omitted.
[2] For code numbers see table 9.19.

The first correlation matrix (table 9.15) is based upon 202 unclassified aviation students, and the second matrix (table 9.18) upon 170 classified pilots. In spite of their classification, the range of ability for these pilots was probably not significantly restricted except perhaps on the Complex Coordination test, which shows slightly reduced factor loadings in this sample as compared with the first.

The two sets of centroid loadings and communalities are given in tables 9.16 and 9.19 and the rotated loadings in tables 9.17 and 9.20.

The Factors

Since most factors presented here are found in both analyses, parallel results will be given. Only loadings regarded as probably significant will usually be mentioned. The criterion of significance is arbitrarily taken to be loadings above 0.20 in both analyses, or above 0.25 in at least one analysis.

Rotated factor I is defined by the following data:

Test numbers		Test name	Loadings	
I	II		I	II
15	2	Spatial Orientation I	0.69	0.64
14	1	Speed of Identification	.65	.65
8	—	Pursuit	.52	[1]....
7	—	Planning A Circuit	.41	....
—	13	Map Planning	....	.45

[1] A dash in these tables indicates the fact that this test was not present in this analysis.

This is clearly the perceptual-speed factor which always comes out clearly when the first two tests in the list are present in the same analysis, and the loadings in those two tests are very stable. The presence here of the Pursuit test and Planning a Circuit test with such strong loadings is a little surprising and gives reason to modify former conceptions of this factor. The two tests are clearly similar to Map Planning in that all of them involve perception of maze-like patterns. Clarity of visual form may consequently have to be added as an aspect of this factor.

Rotated factor II is defined by the following data:

Test numbers		Test name	Loadings	
I	II		I	II
—	6	Numerical Operations	....	0.63
13	5	Mathematics B	0.48	.65
3	—	Organizational Planning	.41	....
4	—	Sequence of Maneuvers	.30	....
14	2	Spatial Orientation I	.24	.27
—	11	Planning A Course	....	.28
—	19	Practical Estimations II	....	.28
—	3	Technical Vocabulary (N)	....	.28

This is the numerical factor. It is interesting to see how this factor creeps into a variety of tests. Organizational Planning involves numbers only as symbols of stations in a map. The stations are numbered systematically, so it is possible that arithmetical computations could have entered into the solution of the most economical paths. Sequence of Maneuvers involves frequent arithmetical problems in computation of altitude changes. Technical Vocabulary (N) probably reflects the numerical factor indirectly due to its coverage of mathematical interest and training. It is not so easy to see numerical work involved in the other tests except in the coding of items and responses.

Rotated factor III is defined by the following data:

Test numbers		Test name	Loadings	
I	II		I	II
13	5	Mathematics B	0.48	0.50
5	—	Spatial Reasoning	.44	....
6	10	Planning Air Maneuvers	.30	.12
3	—	Organizational Planning	.29	....
—	14	Competitive Planning	....	.29
12	4	Reading Comprehension	.23	.19
—	17	Practical Judgment II	....	.40
—	13	Map Planning	....	.31
—	12	Route Planning	....	.27
4	—	Sequence of Maneuvers	.21	....

This is a general-reasoning factor consistently strong in Mathematics B (Arithmetic Reasoning). It is called general because it is common to more tests than either of two other factors that are peculiar to reasoning tests. It can be seen that most of the planning tests have some small but probably significant loadings in this factor.

Rotated factor IV is defined by the following data:

Test numbers		Test name	Loadings	
I	II		I	II
16	9	Complex Coordination	0.56	0.48
5	—	Spatial Reasoning	.33	....
3	—	Organizational Planning	.29	....
7	—	Planning A Circuit	.28	....
6	10	Planning Air Maneuvers	.23	.28
—	11	Planning A Course	....	.62
—	13	Map Planning	....	.27

This is the factor frequently found with stable loadings in the Complex Coordination test and is called spatial relations. It is found with greatest loadings in tests in which either the stimuli or responses have spatial arrangements—right-left, up-down, forward-backward—or both. Other tests strongly loaded with it are the Discrimination Reaction Time test and the Two-Hand Coordination test (see Report No. 4 for description of these tests). The loading of 0.62 in Planning a Course is probably spuriously high since in another analysis the same loading is only 0.34 (see p. 224). Discrepancies as large as this are rare in factorial results. It can possibly be attributed to sampling errors.

Rotated factor V is defined by the following data. Nonsignificant loadings are reported for this factor in planning tests because this factor is of special interest in that connection.

Test numbers		Test name	Loadings	
I	II		I	II
11	8	Mechanical Principles	0.50	0.45
9	—	Driving Skill	.42	....
2	—	Non-mechanical Judgment	.30	....
16	9	Complex Coordination	.28	.09
—	14	Competitive Planning	....	.32
—	17	Practical Judgment II	....	.29
—	13	Map Planning	....	.28
—	12	Route Planning	....	.24
—	11	Planning A Course	....	[illegible]
6	10	Planning Air Maneuvers	.10	−.04
7	—	Planning A Circuit	.03	....
3	—	Organizational Planning	.00	....
4	—	Sequence of Maneuvers	.00	....

This is the visualization factor which apparently entails the manipulation of visual symbols. One might expect planning of various kinds to depend heavily upon some type of visualization, but except for small loadings in Competitive Planning and three other planning tests, this seems not to be true. This may be taken to mean that this factor merely involves a very simple transformation of some perceived or imagined pattern. It apparently does not serve in creative thinking but does seem to help arrive at facts.

Rotated factor VI is defined by the following data:

Test numbers		Test name	Loadings	
I	II		I	II
12	4	Reading Comprehension	0.65	0.69
4	—	Sequence of Maneuvers	.39	....
5	—	Spatial Reasoning	.34	....
13	5	Mathematics B	.32	.27
6	10	Planning Air Maneuvers	.26	.26
11	8	Mechanical Principles	.20	.26
—	15	Practical Judgment I	....	.30

This is the very well known verbal factor. Of all planning tests it is found to a moderate degree only in the Sequence of Maneuvers test. This test is distinct among planning tests for its unusually long and involved verbal instructions. A similar explanation cannot well be given for the loading of 0.26 in Planning Air Maneuvers, however, for its instructions are fairly simple and straightforward. Verbal comprehension must, therefore, enter into these two tests in some other manner, or the loading in Planning Air Maneuvers is perhaps spurious.

Rotated factor VII is defined by the following data:

Test numbers		Test name	Loadings	
I	II		I	II
10	7	Mechanical Information	0.77	0.74
11	8	Mechanical Principles	.61	.64
1	—	Mechanical Judgment	.54	....
9	—	Driving Skill	.46	....
12	4	Reading Comprehension	.39	.49
16	9	Complex Coordination	[illegible]	.30
6	10	Planning Air Maneuvers	.17	.31
—	17	Practical Judgment II	....	.32
—	18	Practical Estimations I	....	.33
—	19	Practical Estimations II	....	.32

The mechanical-experience factor is here shown to be an element in only one planning test, Planning Air Maneuvers, and even in that it is rather weak. In all others, its variance is zero or insignificant.

Rotated factor VIII is defined by the following data:

Test numbers I	Test numbers II	Test name	Loadings I	Loadings II
2	—	Nonmechanical Judgment	0.39	....
4	—	Sequence of Maneuvers	.38	....
1	—	Mechanical Judgment	.36	....
—	17	Practical Judgment II	....	0.45
—	14	Competitive Planning	....	.36
—	18	Practical Estimations I	....	.36
—	16	Practical Judgment I	....	.30

With the four judgment tests all having equivalent loadings on this factor, although no test is in both batteries, the identity of the factors in the two analyses could hardly be mistaken. As a matter of fact, many items in the mechanical and nonmechanical judgment tests in the first analysis were identical with items in Judgment I and II in the second analysis. The best hypothesis for this factor is judgment—the ability to weigh solutions and select the wisest and best one. This interpretation fits Sequence of Maneuvers and Competitive Planning very well. Why Practical Estimations II is not present in the list is unexplainable. This factor was also found in the analysis of judgment and reasoning tests (see p. 152).

Rotated factor IX is defined by the following data:

Test numbers I	Test numbers II	Test name	Loadings I	Loadings II
6	10	Planning Air Maneuvers	0.51	0.46
7	—	Planning A Circuit	.40	....
2	—	Nonmechanical Judgment	.36	....
4	—	Sequence of Maneuvers	.35	....
—	18	Practical Estimations I	....	.47
—	19	Practical Estimations II	....	.36
—	15	Judgment of Proportions	....	.31
—	16	Practical Judgment I	....	.30
8	—	Pursuit	.27	....

The only common ties for these two lists of loadings are the ones for Planning Air Maneuvers and for the two judgment tests. To call these separate factors would inflate the over-all communality of Planning Air Maneuvers to an untenable level. Its communality is 0.71 in the second analysis, which comes very close to its estimated reliability (0.73).

This factor cannot be satisfactorily interpreted at present. One hypothesis might be that it represents an ideational fluency; the man who can think of more solutions per unit of time would have an advantage in some of these tests. Planning a Circuit does not fit this hypothesis very well, however, nor do some of the estimation tests. Another hypothesis might be that this is some form of visualization different from the manipulation type (factor V). This idea fits most of the tests but lacks fully convincing evidence. Why other planning tests do not also require the same type of visualization is hard to understand. Only further data will more clearly define this factor. It had best be left with the general name of planning until more definitive evidence is available.

Rotated factor X appears only in the second analysis:

Test numbers		Test name	Loadings	
I	II		I	II
—	10	Planning Air Maneuvers	..	0.44
—	12	Route Planning	..	.39
—	14	Competitive Planning	..	.33
—	11	Planning A Course	..	.33

This seems to be identifiable with the factor called integration III, to be described in chapter 10.

It is curious that the only factor that has so many of the planning tests in common, and only planning tests to any significant degree, should not be called "planning," that name being given to another factor. In the integration-battery analysis at least, five nonplanning tests have apparently significant loadings on the same factor. It might, after all, be a second planning factor. The term "integration III" merely arises from the fact that it was discovered in the integration battery.

General Conclusions

In conclusion it can be said of planning tests that their fundamental variances break down along different lines, some assignable to already well-known factors, and some to new unidentified factors. No planning test in the list was found to be satisfactorily pure. Most of them have significant, though rather small, loadings in general reasoning. Given in order of their loadings in this factor they are: Map Planning, Planning Air Maneuvers, Competitive Planning, Organizational Planning, Route Planning, and Sequence of Maneuvers. None of them can be recommended, however, as a general reasoning test. All except Organizational Planning and Map Planning have probably significant loadings in the unidentified factor, which may be ideational fluency or a creative visualization rather than planning as such. The strongest tests in this factor are, in order: Planning Air Maneuvers, Planning A Circuit, and Sequence of Maneuvers. Map Planning and Planning A Circuit have strong variances in perceptual speed. Organizational Planning, Sequence of Maneuvers, and Planning A Course have moderate to low loadings in numerical facility. As a negative conclusion, it can be said that planning tests are not mechanical, not visualization tests (of the manipulatory variety), nor are they verbal (except for Sequence of Maneuvers).

When factor loadings are considered in connection with arbitrary groupings of this chapter, it will be seen that there is not much supporting evidence for that type of categorization. Two of the three economical-procedures tests are leaders with variance in the new planning factor. The third, however, is decidedly missing from the list and a test not in the list—Planning A Circuit—is prominent. We cannot, therefore, call the new variable an economical-procedures factor.

All in all, these analyses have failed to demonstrate a clear-cut fundamental ability that should be called either foresight or planning, and while two new interesting factors have been uncovered, no relatively pure test for either of them has as yet been found.

CHAPTER TEN

Integration Tests[1]

INTRODUCTION

Job-Analysis Findings

Inability to pay attention to numerous conditions while engaged in some phase of flying activity and to construct an integrated impression of these conditions quickly and appropriately seemed to be the common element of a variety of stated reasons for eliminations in primary pilot schools. "Unable to think of more than one thing at a time," "frequently becomes confused," "suffers from indecision," "cannot divide attention" —these are typical comments made by instructors regarding failing students who are probably deficient in the ability to integrate.

In a faculty-board account of reasons for elimination of 102 bombardier students, the lack of "ability to execute a series of activities accurately and in proper order" was mentioned as a deficiency in 70 percent of the cases by instructors and check-flight bombardiers.[2]

Analysis of the performance required of the pilot trainee in primary school, or of the navigator or bombardier in advanced school, suggests that the successful air-crew member must maintain sets to respond to a large number of conditions, cues, and reference points. Often these conditions must be observed simultaneously or at least within a brief period of time. Moreover, they must often be noted while some other activity is being carried out, thus making it necessary to divide attention without disrupting the pattern of action in progress. Some cues, when they occur, call for immediate action; others call for delayed responses with which there must be no interference by intervening activities. In order to respond appropriately, the various conditions that influence action at a given time must be observed, remembered, and integrated.

An illustration of these requirements is seen in the pilot's choice of fields during forced-landing practice. The pilot must make most of his observations while establishing and maintaining the proper glide. Certain conditions that will determine the field chosen must be noted quickly while other conditions must be remembered from previous observations. Among the numerous things requiring consideration are (1) direction of the wind; (2) altitude of the plane; (3) relative distances from available fields; (4) surface characteristics of available fields; and (5) hazards to the approach of available fields. Students able to make an

[1] Written by T/Sgt. Sanford J. Mock and the editor.
[2] See table 1.1.

appropriate judgment in relation to any one of the above conditions presented singly may find it very difficult to make a successful integrated response in the presence of a number of them.

Requirements of an Integration Test

In constructing a test for this supposed function, the following considerations were observed:

a. The difficulty of learning any material should be kept at a minimum. If possible, the test score should be unaffected by differences in learning ability.

b. The signs to which the examinee responds should be presented (1) preferably during the conduct of some activity and (2) in such a manner that a number of them must be carried in mind simultaneously.

c. The multiple cues should not be such that they singly lead to separate actions. Rather, they should require integration and the selection of an appropriate response or series of responses governed jointly by the several cues.

In order to fulfill these requirements, the test, in addition, should be built around some common pattern of activity that would be modified in various ways by the test conditions.

An Hypothesis Based on Factor Analysis

The technique of factor analysis contributed another reason for the development of integration tests. The hypothesis was advanced that the "Mashburn factor" or "intellectual component of Complex Coordination" (later identified as spatial relations), which has been a constant component of the Complex Coordination test in all analyses, involved the ability to integrate a number of disparate activities quickly and accurately. In the light of this hypothesis, and for the reasons enumerated previously, a battery of tests was constructed which, it was hoped, would measure the ability to integrate. Tests designated as integration tests were: (1) Planning a Course; (2) Flight Formations; (3) Signal Interpretation; (4) Forced Landings; (5) Combat Planes; (6) Complex Concentration; and (7) Code Analysis.

THE INTEGRATION TESTS

Planning a Course, CI406AX3 [a]

This is the final experimental form of a test in the area of integration. The ability to plan a course of action, considering various factors and exhibiting proper division of attention, is believed to represent one type of integration.

Description.—The examinee learns a simple set of signals to which appropriate responses must be made. He finds these signals scattered

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter, Jr., Capt. Stuart W. Cook, S/Sgt. J. Gordon Elkin, Lt. David H. Jenkins.

through a standard rectangular maze, and he encounters them as he moves through the maze by drawing a line from the beginning to the end of it.

(1) *Internal characteristics.*—This test consists of 15 sample items and 120 scored items. The problems are presented diagrammatically, in series of five items.

(2) *Administration.*—The examinee receives a direction sheet and a work booklet. Answers are marked directly on the work booklet. When the test is finished, answers are transcribed to the regular IBM answer sheet. The time limit for the test is 15 minutes, exclusive of transcription time. In figure 10.1 may be seen the first practice diagram. A part of the directions for the test follow. Administrative directions that are read by the administrator to supplement the directions sheet are printed in italics.

This is a test of your ability to modify a planned course of action. Look at practice diagram I. Your task will be to determine the correct course through similar diagrams. Notice the vertical and horizontal pathways and the entrance marked START. *Also observe the letters, R, L, DS, and CD, which are written at the starting point and above the pathways at various places in the diagram. These letters signal the directions to be taken when the course is traced through the pathways.*

These signals and their meanings are as follows:

R=*One move to the right.*
2R=*Two moves to the right.*
3R=*Three moves to the right.*
L=*One move to the left.*
2L=*Two moves to the left.*
3L=*Three moves to the left.*

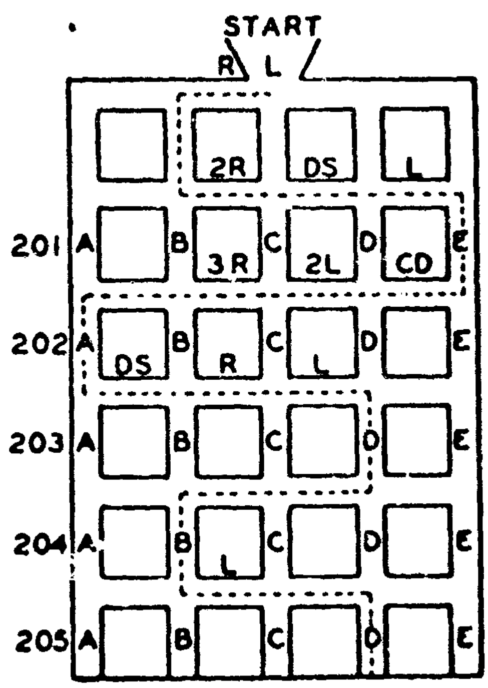


FIGURE 10.1
SAMPLE DIAGRAM OF PLANNING A COURSE,
C 1406AX3

DS=*Double signals. That is, carry out twice all signals which are passed* AFTER *the "DS" signal.*

CD=*Cancel double signal. This signal removes the effect of the DS.*

Beginning at the word start, your task will be to follow the signals in the diagram, tracing a course through the pathways until the bottom line of the diagram is reached. Sometimes you will be moving to the left and sometimes to the right. Before changing the direction of the course from left to right, or from right to left, always make one move down. Never turn back on your course, go down instead.

Now look again at practice diagram I. Use your pencil and trace the course in the diagram as the administrator reads the explanation. As you go through practice diagram I, other rules will be brought out. These rules will be summarized when you are through with the first diagram.

"The course begins at the word *start.* The first move to be made is indicated by the signal which appears under the word start. In this problem the first signal is L. Therefore, the first move is one square to the left *from the starting point.* Make the move now." [Pause.] "This move passes under the signal R, which means "make one move to the right." Because this signal indicates a change in the direction of the course, one move down must be made before moving to the right. This move down passes through column B. Make the down move through column B now." [Pause.] "Now move one square to the right to column C."

"The move to the right passes under the signal 2R, which means 'make two more moves to the right.' Make those moves now." [Pause.] "This carries the course to column E. These last two moves passed under the signals DS and L. As this DS doubles the L signal which follows it, the next move must be twice L, or two moves to the left."

"However, since this is a change in direction, a move down must be made through the letter E before the two moves to the left are made. Now make the move down through E, opposite 201. Now move two squares to the left. This takes you to column C." [Pause.] "These last moves passed under CD and 2L, which means 'Cancel double signal,' and 'Make two moves to the left.' "

"Make these two moves over to column A now." [Pause.] "These moves to the left passed below 3R which, since it calls for a change in direction, will be made after a down move has been made through the letter A, opposite 202. Make the down move through A now." [Pause.] "Now move the three squares to the right to column D." [Pause.]

"In making these three moves to the right, the course passed under a DS, and R, and an L signal, However, since the execution of the 3R signal carried the course to column D, there is not enough space left in this row to move the two squares called for by the doubled R signal; that is, the DS followed by the R. Therefore, the doubled R must be postponed until it can be carried out, and a move down must immediately be made through column D instead. Make the down move through D now, opposite 203." [Pause.]

"The signal L which was also doubled, because it followed DS, also remains to be carried out. Do this now by moving two squares left to column B." [Pause.] "Since there are not more left moves to be made, a down move must be made through B. Make this move through B now, opposite 204." [Pause.]

"The doubled R which has not yet been carried out can now be executed. Make the two moves right to column D now, opposite 205." [Pause.] "Now move down through D since there are no more moves to the right to be made." [Pause.] "Observe that you have reached the bottom of the diagram without being able to carry out the doubled L." [Pause.]

"Note that the cou e passes through a letter whenever a down move is made

opposite one of the numbers on the left. This letter indicates your answer. Do not mark any answer sheet at this time.

Answers will NOT *be recorded until the entire test is completed. The correct answers to practice Diagram I are: 201—E; 202—A; 203—D; 204—B; and 205—D."*

At this point the directions are summarized again, and the examinees work sample problems 2 and 3. Then they are allowed 15 minutes to complete the 120 scored items.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data given below are for classified pilots tested at Psychological Research Unit No. 3 in August 1943.

Distribution statistics.—The distribution of scores in this test is described by a mean score of 71.7 and a standard deviation of 24.3, based on a sample of 877 classified pilots.

(2) *Internal consistency.*—The internal consistency of items in this test is indicated by a mean phi of 0.38, with a range from 0.07 to 0.83 and a standard deviation of 0.14, based on the highest 27 percent and the lowest 27 percent of 800 classified pilots.

(3) *Reliability coefficient.*—A reliability coefficient of 0.81, corrected, was obtained by the alternate-forms method on a sample of 167 classified pilots. This was computed on a preliminary form of the test, which had two parts separately timed. Although the final form of the test was not divided into two parts, the items are sufficiently similar to the previous form so that this reliability coefficient can be considered representative.

(4) *Difficulty.*—The difficulty level of items in this test is indicated by the mean proportion of correct responses equal to 0.72, corrected for chance success. The proportions ranged from 0.16 to 0.99 with a standard deviation of 0.20. These data were based upon results from 167 classified pilots.

(5) *Factorial composition.*—The chief factors are spatial relations (0.45), integration III (0.41), numerical (0.30), and general reasoning (0.24). The communality is 0.64, which is well short of the reliability (0.81).

(6) *Test validity.*—The validity of this test is indicated by a biserial correlation with the graduation-elimination criterion of 0.17, uncorrected. This statistic is based on 877 classified pilots in class 44C. The proportion of graduates was 0.91, the mean score of the graduates was 72.46, the mean score of the eliminees was 64.30, and the standard deviation for all scores was 24.32.

Evaluation.—The test's pilot validity is almost exactly identical with that predicted from its factor pattern (see Table 28.18). It might be revised so as to maximize its spatial-relations loading, in which case it would be one of the best tests available for that factor. Other tests, described in chapter 19, are more promising for this, however. Its numerical variance is no aid to pilot prediction, and its loading in integration III is probably a definite handicap. Its validity for navigator selection promises to be greater than that for pilots, even if its integration III loading is zero in the navigator criterion.

Parenthetically, it is interesting to point out that a test that was developed to function in printed form as the Complex Coordination test does in apparatus form, on the basis of one hypothetical trait (integration), came out strongest in the most valid factor, which in the meantime became recognized as something quite different in character (spatial relations). The spatial characteristics of the task in planning a course had been used as the medium through which integrative aspects of behavior were to be measured. Had the medium been changed, the communality with complex coordination would probably have been lost.

Variations of the test.—Two forms of Planning a Course, CI406AX1 and CI406AX2, preceded the final form. The changes introduced in CI406AX3 were designed to shorten and clarify the directions, although the essential characteristics of the test remained unchanged. Greater simplicity and a more nearly optimum difficulty level were achieved in the final form.

Flight Formations, CI654AX5 [a]

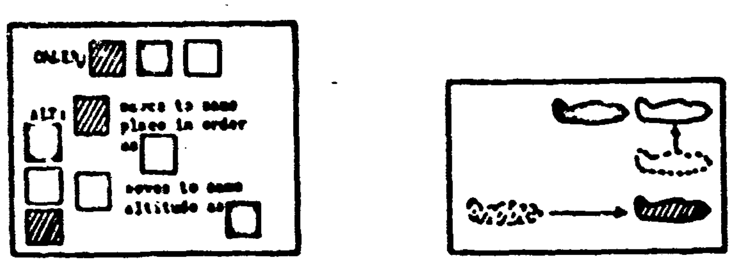
This is the last experimental form of another test in the integration group.

Description.—The examinee must determine the formation of a group of planes after certain moves have been described.

(1) *Internal characteristics.*—The test consists of 1 unrecorded sample problem, 2 recorded but unscored practice problems, and 40 scored items—20 in part I and 20 in part II.

(2) *Administration.*—Twelve minutes are allowed for part I and 10 minutes for part II. Following are the first two pages of directions, and sample problems are given in figures 10.2 and 10.3.

This is a test of your ability to plan ahead of the plane in flight formations. The formation of each flight consists of three planes appearing in different relative positions. Your task will be to determine the new formations of these planes after they have completed certain moves.



A B

FIGURE 10.2
SAMPLE ITEM USED TO EXPLAIN FLIGHT FORMATIONS, CI654AX5

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. William M. Wheeler, Lois G. Wright.

You will be shown the relative order and altitude of the three planes. In working the problems you must first combine this order and altitude into a flight formation. *Order* is the relative positions of the planes from left to right. *Altitude* is the relative positions of the planes from top to bottom.

Look at figure I below. (fig. 10.2a) The small squares represent the three different planes. The squares after the word ORDER show the order of the original formation. Here it is striped plane at the left, dark plane in the center, and white plane at the right. The altitude of the original formation which is shown by the small squares under ALT is: Dark plane at the top, white plane in the middle and striped plane at the bottom. When you combine this order and altitude you should imagine a flight formation with the striped plane at the left and at the bottom, the dark plane at the center and at the top, and the white plane at the right and in the middle. In each problem you must always imagine the original formation by combining the order and altitude before you make any moves. After you have determined the original formation, you must carry out the moves that are described. In figure I, the first move is: Striped plane moves to the same place in order as the white plane. The second move is: White plane moves to the same altitude as the black plane. These moves and the final formation are shown in figure II. (fig. 10.2b)

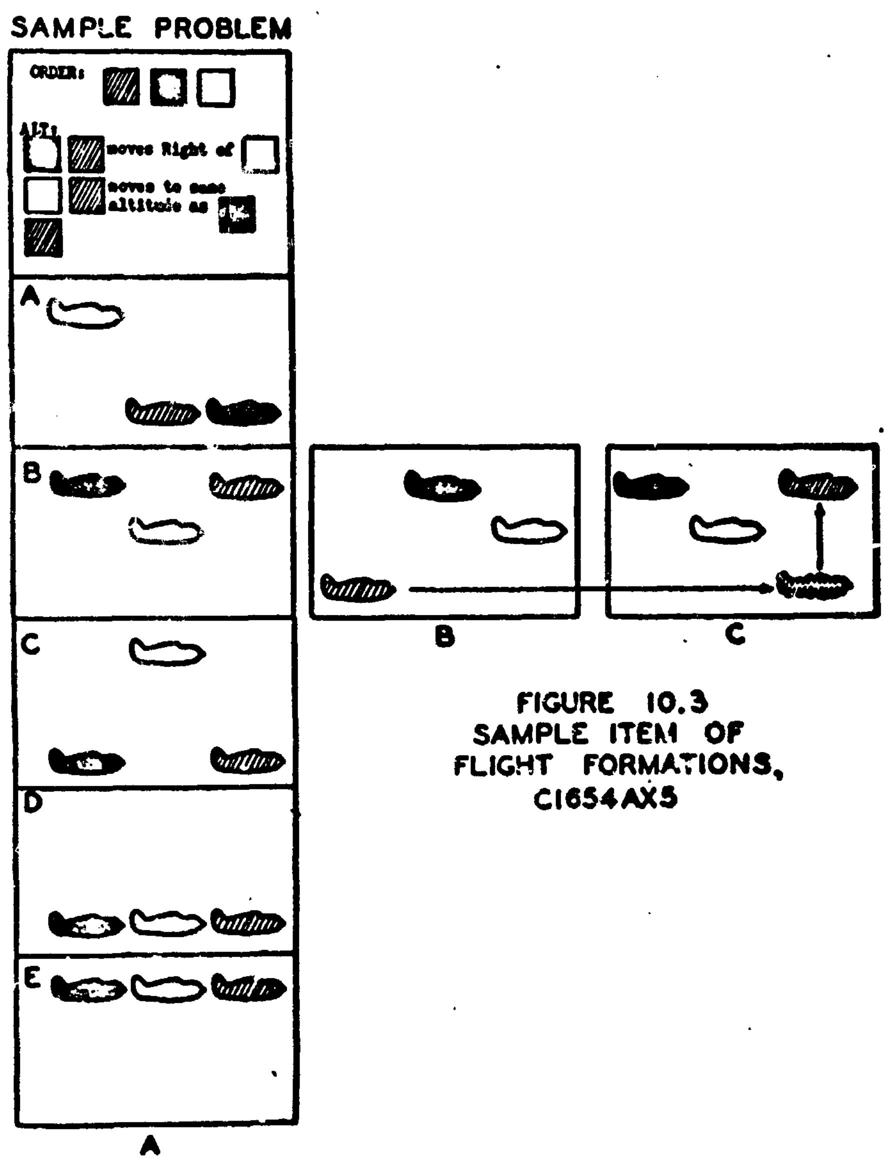


FIGURE 10.3
SAMPLE ITEM OF
FLIGHT FORMATIONS,
C1654AX5

703320—47—14

The dotted lines indicate the original formation positions of the striped and white planes. You must remember that the correct final formation is not determined by the moves alone. The original formation must be visualized before the moves are made.

Now examine the sample problems * * * and imagine what the original formation of the planes should be. (See fig. 10.3a)

If you have interpreted the order and altitude correctly, you should have imagined a formation like the one in figure IV. (fig. 10.3b.)

Keeping the original formation in mind, make the moves called for by the problem. Here you must imagine striped plane moves to the right of white plane, and striped plane then moves to the right of white plane, and striped plane then moves to the same altitude as black plane. Select the correct final formation from the five answers that are given below the problem.

B is the correct answer. Figures IV and V show how the moves should have been made to give you the correct final formation. (See figs. 10.3b and 10.3c)

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—Data for this test are moderately complete, including validations for pilot training. Unclassified aviation students were tested at Psychological Research Unit No. 3 in September 1943, and classified pilots in class 44E were tested at that unit in October 1943.

(1) *Distribution statistics.*—The distribution of scores in this test is indicated by a mean score of 12.0 and a standard deviation of 8.3, based upon a sample of 284 unclassified aviation students.

(2) *Internal consistency.*—The internal consistency of items is indicated by a mean phi of 0.56 for part I, and 0.55 for part II, with a range from 0.05 to 0.95 for the total test and a standard deviation of 0.19 for part I and 0.14 for part II, based upon the highest 27 percent and the lowest 27 percent of 800 unclassified aviation students.

(3) *Reliability coefficient.*—A reliability coefficient of 0.84, corrected, was obtained by the alternate-forms method (Part I *v.* Part II) on a sample of 1,553 classified pilots.

(4) *Difficulty.*—The difficulty level of items in this test is indicated by a mean proportion of correct responses of 0.62, corrected for chance success, a range from 0.41 to 0.99, and a standard deviation of 0.15. These results are based upon the responses of the above-mentioned sample of 800 unclassified aviation students.

(5) *Factorial composition.*—The chief factors are: integration I (0.46), general reasoning (0.22), spatial relations (0.22), and integration III (0.21). The communality is only 0.45, which is to be compared with a reliability of 0.84.

TABLE 10.1.—*Validity data for Flight Formations, CI654AX5, with the graduation-elimination criterion*

Group	N_t	p_s	M_s	M_u	SD_t	r_{bis}	r_{bis}[1]
Pilots in primary training[2]	1,302	0.91	18.30	15.33	7.92	0.19	0.23
Pilots through basic training[2]	1,292	.86	18.39	15.60	7.93	.19	.23

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44E.

(6) *Test validity.*—Data based on a large sample in elementary and basic training are given in table 10.1.

Evaluation.—Flight Formation, CI654AX5, has a moderately high loading in the new factor called integration I. Since its other loadings fail to account for its validity of 0.23, the validity of this factor for pilot selection should be considerable. Assuming that this factor validity is 0.25, the test validity is almost fully accounted for. Because of its unique contribution, this test should be purified. Its reliability is satisfactory. It would have a validity of at least 0.20 for navigators without including any integration variance, whose navigator validity is unknown.

Flight formation might well be expected to be a visualization test. In a factor analysis of the integration battery, however, it revealed a loading of only 0.04 in the visualization factor. No result could be more decisive on this point. The conclusion should be that if this test is one of visualizing, it is a different type than that common to known tests.

Signal Interpretation, CI656A [6]

This is another test in the integration area.

Description.—In each problem there is a diagrammatic representation of 10 airplane carriers in a row. The examinee must determine from which of these carriers planes will take off. This can be ascertained by

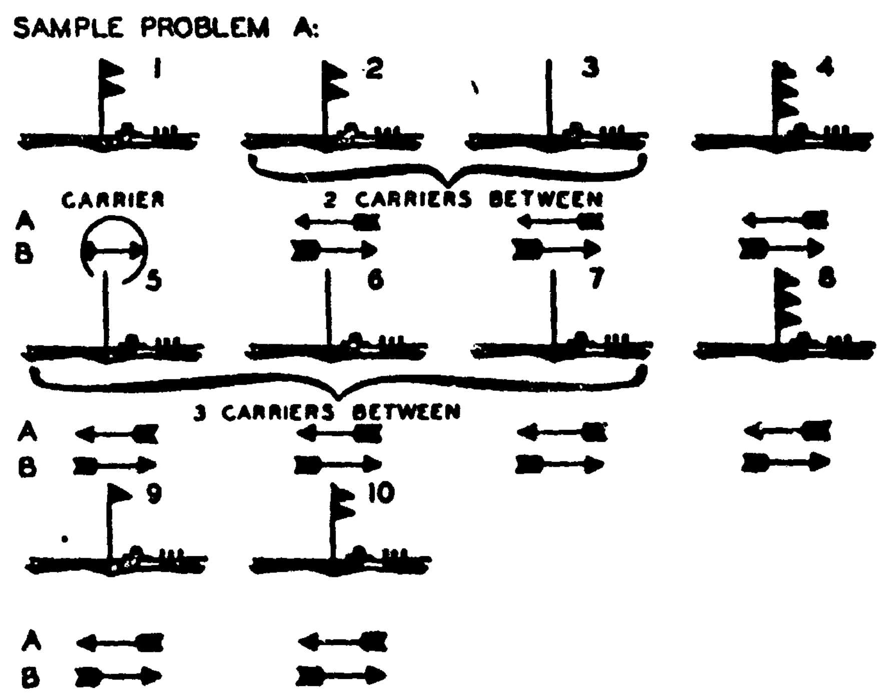


FIGURE 10.4
SAMPLE ITEMS OF SIGNAL INTERPRETATION, CI656A

[6] Developed at Psychological Research Unit No. 3. Chief contributors: S/Sgt. J. Gordon Elkin, S/Sgt. Benjamin Fruchter, and S/Sgt. Wayne S. Zimmerman.

interpreting certain signals such as the number of flags on the mast of each ship, the direction the ship is heading, and whether there are more or fewer flags than on the previous take-off ship.

The examinee must compare quickly the number of flags on each ship with the number on the previous take-off ship. He must decide from the relationship between these two the direction of the next take-off ship.

(1) *Internal characteristics.*—As has been stated, each problem consists of a series of 10 ships. The test consists of two unscored sample problems and 15 scored problems, yielding 150 scored items.

(2) *Administration.*—Answers are marked on an expendable test booklet by circling the appropriate arrow under each item. The examinee also receives a single directions sheet. When the test is completed, he is instructed to transcribe his answers to a regular IBM answer sheet. Testing time is limited to 7 minutes. Approximately 9 minutes are required for administration and 13 minutes for the transcription of answers.

In figure 10.4 is shown the first series of 10 sample items. The following are parts of the directions:

Each problem is made up of a row of 10 ships. Your task is to determine which ships in each row carry planes. You will locate these carriers by following certain signals.

Look at sample problem A on your work booklet. (See fig. 10.4.) Note that some of the ships fly no flags, some fly one flag, some two, and some three. Note also the arrows under the ships. The arrows after the letter A all point to the left; those after B all point to the right. The flags on the ships and the arrows under the ships are the signals which you must interpret in order to locate the aircraft carriers.

Here is the way you must interpret these signals. Follow these directions closely:

1. The first ship in each row is always a carrier. So in the example, ship No. 1 is a carrier.

2. The number of flags on the mast of each carrier shows how many ships lie between it and the next carrier. In the example, the first carrier, ship No. 1, flies two flags. Thus, there are two ships between ship No. 1 and the next carrier. Ship No. 4 is the next carrier.

3. Compare the number of flags on ship No. 4 with the number of flags on the carrier you just left, ship No. 1. You must compare the number of flags in order to determine the direction in which you must go to locate the next carrier ship.

a. If the present carrier flies more flags than the one you just left, you will continue in the direction you are going.

b. If there are the same number or fewer flags on the present carrier than on the one you just left, reverse your direction before looking for the next carrier.

c. You must show your direction by circling the arrow under the present carrier which points in the direction you will go.

As ship No. 4 has more flags than ship No. 1, the previous carrier, circle the arrow under ship No. 4 that points to the right and continue in the same direction in finding the next carrier.

Always be sure to circle one arrow, and only one, under each carrier you locate.

Now complete the first sample problem.

As ship No. 4 flies three flags, and you have circled the arrow under it pointing to the right, you must skip three ships to the right in order to locate your next carrier. This will make ship No. 8 your next carrier. Ship No. 8 flies the same number of flags as the previous carrier, ship No. 4, so you must reverse your direction before locating your next carrier. As your direction is changed, you must circle the arrow under ship No. 8 which points to the left, your new direction.

Ship No. 8 flies three flags. You must now skip three ships to the left to find the next carrier. This is ship No. 4, which had previously been found to be a carrier. When the signals direct you to a ship which you have already marked as a carrier, the problem is completed. Do not circle more than one arrow under any one ship.

(3) *Scoring.*—The last problem is omitted from scoring because of faulty reproduction. The scoring formula is R−W/2. A right response consists in correctly circling or not circling each pair of arrows; the maximum score, therefore, is 126.[4]

Statistical results.—Data on distributions, reliability, and validity are available.

(1) *Distribution statistics.*—Distribution constants for this test are shown in table 10.2.

TABLE 10.2.—*Data on distribution of scores for Signal Interpretation, CI656AX2[1] and CI656A*

Form	Group	N	M	SD
CI656AX2[1] ..	Unclassified aviation students[2]	285	84.3	41.6
CI656A	Classified pilots[3]	1,388	103.5	33.7

[1] A description of this form is given on page —.
[2] Tested at Psychological Research Unit No. 3 on October 1, 1943.
[3] In class 44G. Tested at Psychological Research Unit No. 3.

(2) *Reliability coefficient.*—An odd-even reliability coefficient of 0.77, corrected, was obtained for form AX2 on a sample of 285 unclassified aviation students tested in October 1943 at Psychological Research Unit No. 3.

(3) *Difficulty.*—The difficulty level for form A is indicated by a mean proportion of correct responses of 0.73, corrected, with a standard deviation of 0.20 and a range from 0.16 to 1.00. This result is based upon 727 classified pilots in class 44G who were tested in December 1943 and January 1944 at Psychological Research Unit No. 3.

(4) *Factorial composition.*—The chief factor loadings found in the AX2 form are in the integration I (0.59), general-reasoning (0.41), and integration III (0.30) factors. The communality (0.69) almost reaches the reliability (0.77) but possibly does not account for all of the nonchance variance.

(5) *Test validity.*—The pilot validity of form A of the test is indicated by a corrected biserial correlation of 0.21 with the primary-training graduation-elimination criterion. This statistic is based on 2,112 pilots in classes 44G and 44H who had been tested at Psychological Research Unit No. 3. The percentage of graduates was 0.89, the mean score of the graduates 97.9, the mean score of the eliminees 85.5, and the standard deviation of all was 35.7.

[4] While there are 140 correct responses, each of the 14 scored items has the first correct answer given.

(6) *Item validity.*—The validity of items is indicated by a mean phi of 0.07, a standard deviation of 0.06 and a range from −0.11 to +0.21. This is based on 727 pilots in primary training (class 44G; tested at Psychological Research Unit No. 3), of whom 127 were eliminees.

Evaluation.—Signal Interpretation, CI656AX2, defines, to a greater extent than any other test, a new factor identified as integration I, whose pilot validity has not yet been established but which appears to be near 0.25. The test has a loading of 0.59 in this new factor. Because the test does not have high loadings on known valid factors for pilots, it is reasonable to suppose that it derives much of its validity for pilots (0.21) from the integration I factor. Further experimentation on this factor, therefore, is warranted. The loadings on integration III and general reasoning should by all means be reduced, leaving the test practically pure for integration I.

Variations of the test.—Signal Interpretation, CI656A, was preceded by CI656AX2. This earlier form contained all the essential elements of the A form. Its directions, however, were considerably longer and more complicated, and it was divided into two parts, the task of part II being complicated by additional signals. Each part includes 9 practice items and 90 test items.

This test exemplifies the difficulty encountered in writing directions for all the tests in the integration area. The rationale for the area lists the considerations that were observed in constructing tests to measure integration. From these points it is clear that the tests necessarily had to be complex. Consequently, there was an inherent problem of writing effective directions to describe a complex task.

A revision of the A form, CI656B, was begun but never completed. It represented an attempt to simplify the directions further. A form, CI656C, was later prepared to administer for intercorrelation studies.

Forced Landings, CI652AX4–CI652A [1]

These are the final two forms of another test in the integration group. It was believed that this test would require decisions such as the pilot must make in complex situations involved in forced landings.

Description.—Planes of varying size (single-engine, twin-engine, four-engine) are presented on two-dimensional grids representing (1) altitude and (2) distances to various landing fields which differ in desirability. Wind arrows indicate updraft and downdraft, which add to or subtract from the gliding range. The examinee is required to select for each plane, in order, the best landing field within range. Thus, he must, as quickly as possible, integrate the facts concerning the type of plane, its location in relation to a landing field, effect of the wind on its gliding range, and the desirability of the landing fields in determining the best field upon which to land each plane.

[1] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter, Jr., S/Sgt. J. Gordon Eikin, Pfc. Carl Saxe.

(1) *Internal characteristics.*—Forced Landings, CI652AX4, consists of 3 parts of 30 scored items each. It contains five recorded but unscored sample items. The parts are progressively more difficult. Single-engine airplanes appear in part I. Twin-engine, as well as single-engine airplanes, appear in part II. Twin-engine planes can glide twice as far, winds affect them twice as much, etc. Four-engine, as well as single-engine and twin-engine planes, appear in part III. Four-engine planes can glide three times as far as single-engine planes, winds affect them three times as much, etc. Form CI652A consists of 2 parts of 30 scored items each, and 5 recorded but unscored sample items. Part I contains single-engine planes. Part II contains single- and twin-engine planes.

(2) *Administration.*—The CI652AX4 form was administered as part of the experimental integration battery in September 1943. Then, directions were simplified and the 30 items in part III were dropped in making the CI652A form. It was decided to delete part III, because the test was lengthy, and the interpart correlations were high.

Eight minutes are allowed for part I, 7 minutes for part II, and (in the CI652AX4 form) 7 minutes for part III. A single page of directions is provided. Following are the directions for part I of CI652A; the five sample problems are reproduced in figure 10.5. Comments made by the administrator, which are not on the directions sheet, are italicized.

This is a test of your ability to make decisions quickly in problems similar to forced landings.

The planes are numbered consecutively, while the fields are lettered A, B, C, D, or E. Each plane is directly above the field on the same line. The diagrams show only two dimensions; altitude and horizontal distance.

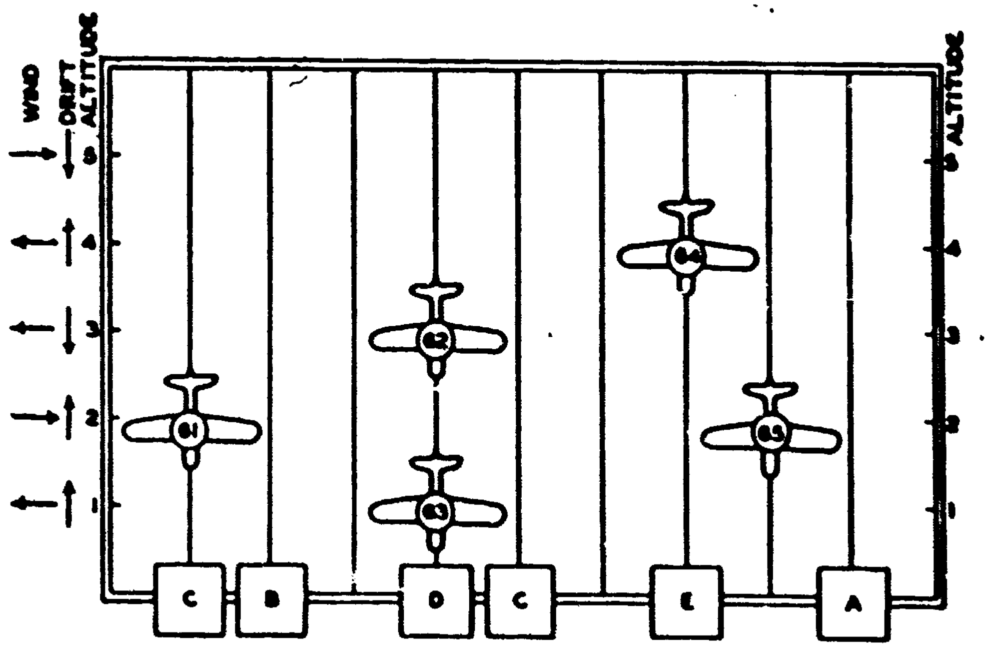


FIGURE 10.5
SAMPLE PROBLEMS OF FORCED LANDINGS, CI652A

a. Altitude, in miles, is shown by the numbers at the left and right of the diagram.

b. The distance from one vertical line to the next is one mile. The distance a plane needs to glide to get to a field is measured sideways only. Count the spaces between the plane and the field to determine the number of miles between them.

Each arrow at the left indicates the direction of winds affecting any planes at the same altitude.

Now look at the front of the work booklet. Hold it up in front of you.

The planes are in the air directly above the field on the same line. For example, plane 61 is directly above the field C. The higher the planes are on the chart, the more altitude they have. For example, plane 61 is 2 miles high as indicated by the number at the left. How high is plane 62? [Pause.] *Three miles is right.*

How far a plane needs to glide to get to any field is measured by ground distance only. For example, plane 61 is 1 mile from field B. How far will plane 61 glide to get to field E? [*Pause.*] *Six miles is correct, because we ignore altitude in computing distance. How far is plane 62 from field A?* [Pause.] *Five miles is correct.*

Follow along on your instruction sheet again. Glance at your diagram from time to time.

Gliding range:

A plane can glide in either direction as many miles as it has miles of altitude. The arrows at the same altitude from which a p[illegible]e starts indicate the winds which will affect the gliding range of the plane. Their effect is as follows:

A. Gliding with the wind adds 1 mile to your gliding range.
B. Gliding against the wind subtracts 1 mile from your gliding range.
C. An updraft (↑) adds 1 mile to your gliding range.
D. A downdraft (↓) subtracts 1 mile from your gliding range.

A plane is not affected by any winds or drafts other than those at the altitude from which it starts.

Landing rules:

A. Do not land a plane on the field over which it starts.

B. Land the planes in order (first No. 1, then No. 2, then No. 3, etc.).

C. Two planes may not land in succession on the same field. For example, plane 2 may not land at the same field, but 1 and 3 may.

D. Land at the best available field. The fields are graded A, B, C, D, and E. Field A is best. Field E is worst.

E. Land at the nearer of two equally good fields. For example, if a plane has a choice of two grade C fields, choose the nearer one.

F. Measure the distance between a plane and the landing field by counting the spaces only, not the altitude.

Caution.—Once you have chosen the best field for a plane, do not go back and change it to get a better grade field for the next plane.

Now look at the front of your work booklet, below the diagram.

Here we have the steps to follow in landing plane No. 61. Notice on the diagram the altitude as indicated by the numbers, and the wind and updraft as indicated by the arrows, as we figure the gliding range in each direction for plane No. 61.

Step One.—Gliding range to your left is 2 miles (altitude gives 2 miles, updraft adds 1 mile, and going against the wind subtracts a mile). Gliding range to your right is 4 m[illegible]ude gives 2 miles, updraft adds 1 mile, and going with the wind adds a [illegible]e).

Step Two.—Note the fields within gliding range. Fields B, D, and C are within 4 miles on your right, none is to the left. The other C field cannot be used because plane No. 61 is directly above it.

STEP THREE.—Select the best field. Field B is best, because B is a better grade field than C or D.

On your answer sheet, black in the space under letter B after item 61.

Notice that plane 61 went to a field only 1 mile away, even though it could have glided 4 miles. In other words, you can land a plane at fields anywhere within the gliding range.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data below are for classified pilots in class 44J tested at Psychological Research Unit No. 3 from April 10 to 13, 1944, and for unclassified aviation students tested at that unit on March 9 and April 16, 1944, and on August 27, 1943.

(1) *Distribution statistics.*—Typical distribution constants for this test are shown in table 10.3.

TABLE 10.3.—*Distribution constants for Forced Landings, CI652A and CI652AX4*

Form	Content	Group	Score	N	M	SD
A	Two parts	Classified pilots	R	540	26.8	12.2
AX4 ...	Three parts	Unclassified aviation students	R−W/4	284	36.4	20.1

(2) *Internal consistency.*—The internal consistency of items in form CI652A is indicated by a mean phi of 0.41, with a range from 0.10 to 0.70 and a standard deviation of 0.15, based on the highest 27 percent and the lowest 27 percent of 750 classified pilots.

(3) *Reliability coefficient.*—This has been estimated from two samples by the alternate-forms method. The data are presented in table 10.4.

TABLE 10.4.—*Reliability coefficients for Forced Landings, CI652A and CI652AX4*

Group	Form	Parts	N	r'_{tt}	r_{tt}
Classified pilots	CI652A ..	Part I v. Part II (right answers only)	1,378	0.65	0.79
Unclassified aviation students .	CI652AX4	Part I v. Part II	284	.76	.91
		Part I v. Part III	284	.71	.88
		Part II v. Part III	284	.67	.86

(4) *Difficulty.*—The difficulty level of items in form CI652A is indicated by the mean proportion of correct responses equal to 0.47, corrected for chance success with a standard deviation of 0.22 and a range from 0.18 to 0.79 based on a sample of 750 classified pilots.

(5) *Factorial composition.*—The chief factors of form CI652AX4 are general reasoning (0.53) and integration II (0.38). No other loading exceeds 0.18 (verbal). The communality is 0.53, which is considerably short of the reliability.

(6) *Test validity.*—Validation has been determined for part scores and total score, also right scores and wrong scores. The data are presented in table 10.5.

TABLE 10.5.—*Validity data for Forced Landings, CI652A, based on pilots in primary training,*[1] *with the graduation-elimination criterion*

Part	Score	N_t	p_s	M_s	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]
I	Rights	1,310	0.90	17.03	16.46	6.40	0.05	0.10
I	Wrongs ...	1,310	.90	4.73	4.52	4.49	.02	.01
II	Rights	1,310	.90	15.12	14.94	7.34	.01	.06
II	Wrongs ...	1,310	.90	5.17	4.94	4.08	.03	.01

[1] In class 44J.
[2] Assuming an unrestricted stanine standard deviation of 2.00.

Evaluation.—Forced Landings, CI652AX4, is a fair test of general reasoning. Its validity for pilots should be only about 0.12, allowing for some validity for the integration factor. The navigator validity would probably exceed 0.20 from the reasoning component alone. Because of the intrinsic complexity of the task in forced landings, it is doubtful that the reasoning loading could be decreased if that were desired. If it could be rid of its loading in integration II, it would be the best general reasoning test developed in the program. Possibly this test could be developed as a navigator-selection instrument, but it lacks promise as a pilot test.

Combat Planes, CI655AX5 [6]

This test emphasizes the ability to carry out complicated directions, keeping in mind restricting rules.

Description.—In this test two squadrons of planes in mock combat are represented. The planes vary in type (single-engine, twin-engine, and four-engine). The task is to determine as quickly as possible, from mock combat rules given in the directions, which opponents each plane can attack. The examinee is required to indicate which opponents are attacked and (in parts II and III) whether a plane stays on the offensive or changes to the defensive. In making these decisions, he must take into consideration the size of each plane, the proximity of the opponents, and the identity of the squadron which starts the offensive.

(1) *Internal characteristics*.—Part I of Combat Planes, CI655AX5, contains 1 sample column of 6 unrecorded and unscored items, 1 practice column of 10 recorded but unscored sample items, and 90 scored items. Part II contains a sample column of 6 unrecorded and unscored items, 1 practice column of 10 recorded but unscored sample items, and 60 scored items. Part III contains 90 scored items.

(2) *Administration*.—Four and one-half minutes are allowed for each part.

Following are the directions and sample items for part I. The words in italics are part of the administrative directions and do not appear in the test booklet.

[6] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Sanford J. Mock and S/Sgt. Wayne S. Zimmerman.

In this test you will be asked to apply rules to a simplified serial combat situation. You will have to determine:

1. If a plane can attack, and
2. Whom it can attack.

In order for a plane to attack, it must meet three conditions:

Condition 1. It must be on the offensive.

Condition 2. It must be immediately next to one or two opponents.

Condition 3. It must be the same size or smaller than its opponents.

Look at the diagram below. (See fig. 10.6.)

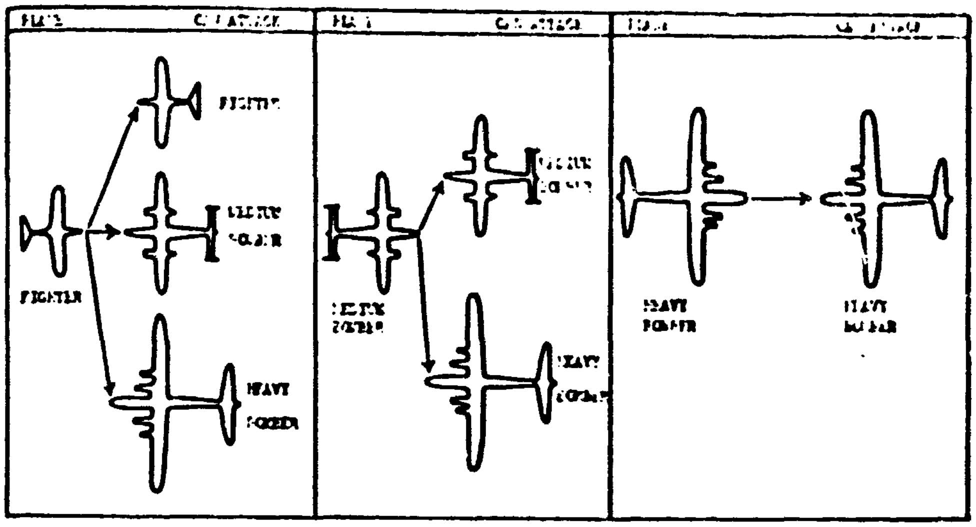


FIGURE 10.6
EXPLANATORY DIAGRAMS FOR COMBAT PLANES,
C1655AX5

Observe that fighters (single-engine) can attack any type of plane—other fighters, medium bombers, or heavy bombers. Medium bombers (twin-engine) can attack the medium bombers and heavy bombers. Heavy bombers (four-engine) can attack only other heavy bombers.

At the top of each column you will be told whether White or Black Squadron is on the offensive. Sometimes it will be one, sometimes the other. Be sure to check this for each column.

In the sample, Black Squadron is on the offensive. (See fig. 10.7.)

Plane S1 is on the offensive; it is next to an opponent, and it is smaller than its opponent. Therefore, it meets the three conditions and can attack plane S2 below it. Plane S2 is on the defensive. It does not meet condition 1, and therefore, cannot attack. Plane S3 is on the offensive and is next to two opponents. Is it the same size or smaller than its opponents? [Pause.] *It is larger than plane S2 above; therefore, it cannot attack this plane. It is the same size as plane S4 below, and therefore, can attack S4. S4 is on the defensive and cannot attack. S5 is on the offensive, is next to an opponent, and is smaller than its opponent. It can attack plane S4 above. Plane S6 is on the offensive but is not immediately next to one or two opponents. Therefore, it cannot attack.*

RULES FOR MARKING ANSWERS

Mark A on your answer sheet opposite the number of the problem, if the plane above is attacked. A stands for above.

Mark B on your answer sheet opposite the number of the problem, if the plane below is attacked. B stands for below.

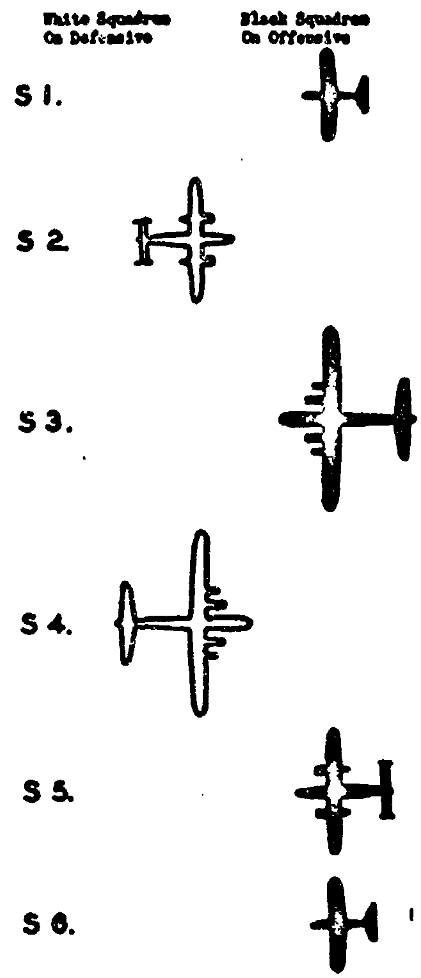


FIGURE 10.7
SAMPLE PROBLEMS FOR COMBAT PLANES,
C1655AX5

Mark C on your answer sheet opposite the number of the problem, if both planes (above and below) are attacked.

Mark D on your answer sheet opposite the number of the first plane that is on the defensive in each column. D indicates that this plane and members of the same squadron below it are on the defensive. Make no marks for any other defensive planes.

Mark E on your answer sheet opposite the number of the problem if the plane is on the offensive, but can not attack.

In parts II and III, a complication is introduced. The new rule is that if a plane has two opponents and cannot attack either of them, the offensive changes. In addition to applying this new rule, the examinee must now not only answer D for the first defensive plane in each column, but also for every plane with which the offensive changes.

Each part of the test is separately timed. The test is highly speeded.

(3) *Scoring.*—The scoring formula is R−W/5.

Statistical results.—Data are quite limited but are sufficient to support a tentative evaluation. The results are for examinees tested in October 1943 at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—For the three parts separately, data are given in table 10.6.

TABLE 10.6.—*Distribution constants for the three parts of Combat Planes, CI655AX5, based on 273 unclassified aviation students*

Part	M	SD
I	20.6	15.3
II	24.6	10.9
III	35.3	15.5

(2) *Reliability coefficients.*—Some indication of reliability of the parts is given in table 10.7.

TABLE 10.7.—*Alternate-forms reliability coefficients for Combat Planes, CI655AX5, based on a sample of 273 unclassified aviation students*

Parts	r'_{tt}	r_{tt}
Part I versus Part II	0.66	[1]...
Part I versus Part III	.65	.85
Part II versus Part III	.81	...

[1] Owing to disparities in part dispersions, the Spearman-Brown correction formula was not applied in all three combination.

(3) *Factorial composition.*—The factorial picture of this test shows noteworthy loadings in the integration I (0.57), general-reasoning (0.33), verbal (0.31), and integration III (0.28) factors.

Evaluation.—Combat Planes, CI655AX5, helps to define a new factor, identified as integration I which accounts for 32 percent of the total variance of the test. Since the test was not administered for validation, nothing positive can be said concerning its validity for air-crew success. Based upon the factor validities given in table 28.17, however, the predicted pilot validity is 0.18. This is equal to the expected validity for Signal Interpretation, which it resembles closely factorially. Conclusions regarding that test also apply here. The chief difference is that Signal Interpretation has a larger reasoning variance and this test a larger verbal component. Both need purifying.

Variations of the test.—Four forms preceded Combat Planes, CI655AX5. Statistical analysis was not responsible for the development of these forms. Indeed, no data are available for the early revisions. The changes were prompted by the necessity for making the directions more readily understandable.

Forms X2, X3, and X4 were known by the title of Attacking Planes. X2 involved a concept of support. Even if a plane could meet the three prescribed conditions, and it faced two opponents, it could not attack unless given support from the nearest plane or planes heading in the same direction. An adjacent plane could give support only if both the opponents of the attacking plane were of the type the supporting plane could attack. This concept proved too difficult to explain adequately in a directions period of limited duration. The task in the X3 revision was simplified by omission of this support concept until part II. This gave the examinees an opportunity to become familiar with the basic problems of the test in part I before encountering the complex matter of support. This also proved unsatisfactory, however, and all references to support were dropped in the X4 form. Instead, the idea of a change in offensive was introduced in part II. If a plane had two opponents and could not attack either of them, the offensive changed to the other squadron. Another important change in form X4 was the placement of the items in vertical columns. In forms X2 and X3 the offensive and defensive squadrons were opposite each other in horizontal rows. The vertical presentation of items facilitated the administration of the test. This feature was incorporated in the final form of the test, Combat Planes, CI655AX5, together with numerous improvements in the arrangement and writing of the directions.

Complex Concentration, CI658AX1 [a]

This is the only test, in the area of integration, in which motion pictures were used. There were two factors that determined the presentation of the test in motion-picture form. First, the test involved the use of color, and it was extremely difficult to obtain color printing. Another circumstance determining the use of motion pictures was the precise timing required. Most of the color patterns were to be exhibited for only three seconds. Uniformity in each administration would have been extremely difficult to achieve if test booklets and ordinary testing procedures had been used.

The material to be photographed, which consisted of varying numbers of 2-inch colored squares, was mounted on a set of gray background cards. Directions were printed and mounted on similar cards. These were placed in proper order, then each card was photographed separately, being exposed to the camera for a predetermined period. The result is a relatively smooth sequence of contiguous, immobile cards.

Description.—For each problem, three groups of differently colored squares are presented on the screen in rapid succession. Each group contains three colors and is visible for three seconds. At the conclusion of each series of three groups, the examinee must record the total number of times he believes each color appeared. Thus, as he sees each new group within a set, he must add the colors to his previous totals in that set without forgetting or confusing colors or frequencies. As the test progresses, the number of colors included in each group increases from three to four and then to five.

(1) *Internal characteristics.*—The test consists of two unscored sample series. The test is divided into two parts. Each part contains 16 series

[a] Developed at Psychological Research Unit No. 3. Chief contributor: Sgt. Hyman Heller.

and 61 scored items. Running time for the film is 27 minutes. Transcription of answers from the work sheet to the regular 5-place IBM answer sheet requires approximately 10 minutes.

(2) *Administration.*—Each examinee receives an expendable work sheet on which answers are first recorded. Sufficient lighting is provided in the test room to allow the examinee to see his work sheet, without radically decreasing the visibility of the image on the screen. Each examinee is given a pencil with the eraser removed. This is to prevent tallying of colors each time they appear, on the theory that the examinee will not do so if he cannot erase his tally marks.

Although the examinees are not led to expect it in advance, erasers are distributed at the beginning of the transcription period so that errors in transcribing answers to the IBM answer sheet can be corrected.

All instructions, except those for transcribing answers, are in film subtitles. There is no sound track.

After preliminary instructions are given, the first sample series appears. It consists of 1 blue and 1 red square, shown for 3 seconds. Then a gray blank background appears for 2 seconds, followed by 2 red squares. Again the gray background, and then two red squares and one blue appear. Now the examinees are instructed to record their answers on their work sheets. While answers are being marked, a gray background appears on the screen for 10 seconds. As the items become more difficult, this answer period is lengthened to 15 and then to 20 seconds.

The answers to the first sample series are graphically illustrated when a hand appears on the screen and writes the answers in the proper places on a work sheet. At the same time, the directions read, "For sample 1, your answers should be 5 red and 2 blue." Another sample series is then presented before part I begins.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data for pilots given below are for examinees in classes 44E and 44F tested at Psychological Research Unit No. 3 in October and November 1943 respectively. The data for navigators are for examinees tested at Selman Field from June 19 to June 22, 1944, and at Psychological Research Unit No. 3 from May 4 through May 6, 1944.

(1) *Distribution statistics.*—The distribution of scores in this test is described by a mean of 62.7 and a standard deviation of 15.9, for parts I and II combined, based on a sample of 856 classified pilots in classes 44E and 44F.

(2) *Internal consistency.*—The internal consistency of items is indicated by a mean phi of 0.29 with a range from 0.06 to 0.51 and a standard deviation of 0.08, based on the highest 27 percent and the lowest 27 percent of 473 classified pilots in class 44F.

(3) *Reliability coefficient.*—A reliability coefficient of 0.77, corrected, was obtained by the alternate-forms method (Part I *v.* Part II) on a

sample of 473 classified pilots in classes 44E and 44F, and of 0.84 based on 668 navigation students.

(4) *Difficulty.*—The difficulty level of items in the test is indicated by the mean proportion of correct responses equal to 0.50, with a standard deviation of 0.18 and a range from 0.12 to 0.94, based on 800 classified pilots in 44F.

(5) *Test validity.*—The test has been validated for both pilot and navigation training, as shown in table 10.8.

TABLE 10.8.—*Validity data for Complex Concentration; CI658AX1 with the graduation-elimination criterion*

Group	Part	Score	N_t	p_s	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training[2]	Total	Formula	381	0.94	63.60	65.70	16.35	−0.06	−0.05
Pilots in primary training[3]	"	"	475	.94	61.86	62.49	15.47	−.02	.04
Navigators	I	Rights	668	.88	43.43	40.96	7.93	.16	.28
Navigators	I	Wrongs	668	.88	17.56	20.16	7.43	−.18	−.31
Navigators	II	Rights	668	.88	46.97	43.75	8.72	.20	.29
Navigators	II	Wrongs	668	.88	14.36	17.64	8.33	−.21	−.31

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44E.
[3] In class 44F.

(6) *Right-wrong correlations.*—For the navigator sample the correlation of rights and wrongs scores within parts I and II were −0.92 and −0.96; between parts the rights-wrongs correlations were −0.68 and −0.69.

Evaluation.—Complex concentration was not factor-analyzed with the integration battery, because it had not been ready at the time the battery was administered. Validity data for pilots are sufficient proof that the test does not offer promise as a selection instrument for that specialty. For navigators a satisfactory level of validity is shown, and because the correlation with the navigator stanine at the time was sufficiently lov (approximately 0.30), the test offered some degree of uniqueness.

Code Analysis, CI653AX3 [10]

This is the final form of another test in the integration area. It was designed as a measure of speed and facility in understanding and analyzing interchangeable symbols and keys such as might be used in a code.

Description.—In each item, a key number series composed of five digits is presented. Below the key series appear five other series, each composed of five digits. These series are the choices from which the examinee is required to select the correct answers. The general task of the examinee is to determine those alternative series that contain the same digits as the key series. Some problems call for the determination of alternatives which contain all the digits found in the key series; others, for alternatives containing four and only four of the digits

[10] Developed at Psychological Research Unit No. 3. Chief contributor: Lt. William M. Wheeler.

found in the key series; or three and only three, etc. Which of these determinations is required is coded within each key series itself, and must be determined by the examinee. Later in the directions, alphabetical substitutes for numbers (A=1, B=2, etc.) are introduced. Thus, the examinee must interchange letters and numbers in solving the codes. He must be aware of the digits (or letters and digits) of the original series, the interchangeability of letters and digits, and the requirements of the particular problem in selecting the proper series as the answers.

(1) *Internal characteristics.*—The test consists of 7 recorded but unscored sample items, and 48 scored items, divided into 2 parts of 24 items each.

(2) *Administration.*—Thirteen minutes are allowed for part I and 11 minutes for part II. Examinees are paced by the administrator who informs them when there are but 6 minutes left to finish part I and also when there are but 5 minutes to finish part II. Following are parts of the directions and sample items. The words in italics are oral administration directions and do not appear in the test booklet.

This is a test of your quickness at understanding and analyzing interchangeable symbols and keys such as may be used in a code.

You will be shown a key number series composed of five digits. Below this will be five other series. These will be the five choices from which you are to select the correct answer series which contains the digits called for by that particular item. Now, work sample 1.

Sample 1. In which of the following series are *all* five of the digits the same as in the key series 9 5 8 7 5 ?

A. 9 5 6 8 7.
B. 9 9 8 7 5.
C. 5 7 8 5 4.
D. 5 6 8 7 2.
E. 5 5 9 8 7.

(Administrator reads the sample)

E is the correct answer. All five of its digits are the same as the digits in the key series. Notice that the digits in the answer series do not have to be in the same order as in the key series. A is not correct, because it does not have two fives, B does not have two fives, C has no nine, D has no nine, and has only one five. Notice that in E the five appears twice, the same as it does in the key series.

Later parts of the instructions read as follows:

Up to this point, in each sample problem you have been told how many characters to look for in the correct answer series. Actually, this information is contained in the key series itself. The first five digits (1, 2, 3, 4, and 5) are code digits. 6, 7, 8, and 9 are not code digits. The first one of the code digits that appears in the key series indicates the number of corresponding digits in the correct answer.

Now look back at sample 1. The first code digit that appears in the key is digit 5. This indicates that all five of the digits have to be in the correct answer. Look at sample 2. The first code digit that appears in the key is the digit 4. This indicates that four and only four of the digits in the key series are found in the correct answer series. Look at sample 3. The first code digit is 3, indicating three of the digits are found in the correct answer. Look at sample 4. The first code digit in the key series is the digit 3. This indicates that three digits have to be found in the correct

703320—47—15

answer. Obviously, 6, 7, 8, and 9 are not used to determine the number of corresponding characters in the answer, since there cannot be more than 5 corresponding characters in a series containing only 5 characters.

* * * Letters as well as digits will be included in the series. The following letters are used—A, B, C, D, E, F, G, H, J. The first five of these letters, i.e., A through E, are code letters. Each of the code letters can be substituted for its corresponding code digit, 1 through 5 respectively. That is, A can be substituted for 1, B can be substituted for 2, C for 3, D for 4, and E for 5. This substitution is reversible; thus, 1 can be substituted for A, 2 for B, 3 for C, etc. The letters, F, G, H, and J are not interchangeable with the digits 6, 7, 8, 9 and no substitution can take place.

In the following problems, the first code digit or its equivalent code letter found in the key series indicates the number of corresponding characters found in the correct answer.

If the key is "JG 4 3 2" the first code digit is four. This indicates that four and only four of the characters in the key appear in the right answer. If the key is "6 C A D 2" the first code letter is C. Since C is equivalent to three, this indicates that three and only three of the characters in the key appear in the right answer. Now, work sample 5.

Sample 5. F E 5 3 A

A. H 3 3 3 A
B. 3 3 5 H 3
C. H C 1 E D
D. F 1 E E 3
E. 1 3 3 H D

D is the right answer. Notice that E is the first code letter found in the key, and since E is interchangeable with five, this indicates that all five of the characters in the key series are found in the correct answer. In choice D, F corresponds to F in the key, 1 corresponds to A, E corresponds to E, and the second E corresponds to five and 3 corresponds to three.

Remember the code characters:

A	B	C	D	E
1	2	3	4	5

F, H, H, J are not interchangeable with anything. Also 6, 7, 8, 9 are not interchangeable with anything.

(3) *Scoring.*—The scoring formula R−W/4 is used.

Statistical results.—The data given are for examinees tested at Psychological Research Unit No. 3 on March 9 and May 16, 1944, and on August 27, 1943.

(1) *Distribution statistics.*—The distribution of scores in this test is indicated by a mean score of 14.9 and a standard deviation of 11.7, based on a sample of 285 unclassified aviation students.

(2) *Reliability coefficient.*—A reliability coefficient of 0.89, corrected, was obtained by the alternate-forms (part I–part II) method on a sample of 285 unclassified aviation students.

(3) *Factorial composition.*—The leading factors and their loadings are: integration III (0.42), integration II (0.40), numerical (0.29), verbal (0.23), and general reasoning (0.20). The communality (0.59) falls far short of the reliability.

Evaluation.—Code Analysis has never been administered for validation. Its factor loadings, however, indicate that it would not possess much validity for pilot success, probably not over 0.07. It has no loading greater than 0.16 in any factor of known pilot validity. If the two integration factors are valid for navigator selection, however, the test would have high navigator validity, for the combination of other factors is very favorable.

A FACTOR ANALYSIS OF INTEGRATION TESTS [11]

In order to gain a better understanding of tests in this area, a special factorial study was made of integration tests. Integration tests were originally predicated on the hypothesis that the most valid aspect of the Complex Coordination test was its measurement of the ability to observe a complicated situation and to make a single integrated response to it. Subsequent findings as related in this chapter have shown that while the hypothesis was in error, it was none the less fruitful in directing research into virgin areas. The newly-discovered territory needs additional illumination through factorial study.

The Data

In addition to Complex Coordination, other classification tests which had been recognized as good factorial reference tests were included in the analysis. These tests have all been described in this volume except the Two-Hand Coordination test. This test uses the familiar lathe-type machine in which the examinee attempts to keep a contact point in touch with a moving button which follows an irregular pathway on the surface of a slowly revolving disc, at irregular speeds. The right-and-left and to-and-fro movements of the contact point are executed independently by turning the cranks of the machine, one in each hand.

In addition to seven experimental integration tests, a number of other experimental tests were also in the battery. The list with code numbers may be seen in table 10.10. It includes some of the planning tests which were of particular interest at the time and some new reasoning and spatial tests. Another hypothesis concerning the "intellectual component of the Complex Coordination test" was that it is a space factor of some kind, hence the inclusion of spatial tests.

Two tests devised for this battery are of special interest—Log Book Accuracy and Marking Accuracy (see ch. 16). It was thought that in many of the integration tests, due to the great amount of rapid, clerical-type work involved, part of which is in the use of the answer sheet, much of the variance would be taken up with some kind of simple psychomotor factor. The two tests were accordingly devised to isolate that hypothetical factor and to determine its possible variance in the integration tests. All the experimental tests included in the integration battery have been described in detail in this volume.

[11] Executed by Capt. Lloyd G. Humphreys, Lt. David H. Jenkins, and S/Sgt. Wayne S. Zimmerman at Psychological Research Unit No. 3.

TABLE 10.9—*Correlation matrix for*

Test	1	2	3	4	5	6	7	8	9	10	11	12
1. Speed of Identification[2]	....	49	07	05	06	07	05	−09	05	14	14	09
2. Spatial Orientation I	49	....	20	26	14	20	20	−07	−02	18	16	27
3. T. V. N.	07	20	....	58	45	23	24	15	19	14	16	28
4. Reading Comprehension	05	26	58	....	52	17	23	14	28	16	30	35
5. Mathematics B	06	14	45	52	....	35	42	10	24	16	29	45
6. Numerical Operations, Front	07	20	23	17	35	....	50	−18	−21	10	−04	29
7. Numerical Operations, Back	05	20	24	23	42	50	....	−03	−02	24	10	39
8. Mechanical Information	−09	−07	15	14	10	−18	−03	....	44	00	16	−05
9. Mechanical Principles	05	−02	19	28	24	−21	−02	44	....	10	25	11
10. SAM Complex Coordination	14	18	14	16	16	10	24	00	10	....	28	23
11. Planning Air Maneuvers	14	16	16	30	29	−04	10	16	25	28	....	37
12. Planning A Course	00	27	28	35	45	20	39	−05	11	23	37	....
13. Instrument Comprehension I	14	30	22	25	28	22	28	−01	09	27	25	36
14. Instrument Comprehension II	22	29	27	29	24	20	17	−01	24	29	20	36
15. Figure Analogies	22	32	35	45	45	22	28	01	24	30	37	46
16. Spatial Visualization I	22	24	37	41	46	11	14	15	40	20	41	43
17. Map Distance	01	06	18	22	23	−00	00	22	22	12	03	07
18. Hands	13	06	01	15	12	09	15	10	18	11	11	14
19. Cubes	34	34	18	15	20	21	23	−08	12	29	32	25
20. Route Planning	19	22	01	28	28	01	10	11	31	18	23	36
21. Organizational Planning	13	25	19	22	23	28	28	07	02	21	29	40
22. Following Oral Directions	16	20	23	33	37	19	27	05	25	22	29	42
23. Following Directions	13	22	27	44	32	24	34	−09	10	19	20	32
24. Code Analysis	11	29	24	42	41	29	34	−09	03	22	32	55
25. Flight Formations	15	23	23	31	31	16	21	−03	13	20	25	38
26. Forced Landings	11	19	22	31	35	10	14	−05	07	14	30	38
27. Signal Interpretation	10	21	26	34	38	17	20	01	15	18	31	41
28. SAM Two-Hand	02	07	−08	−00	−03	−04	03	20	18	33	12	07
29. Combat Planes	08	18	31	46	44	26	35	−03	12	34	33	47
30. Log Book Accuracy	15	27	15	16	14	41	41	−19	−14	16	07	35
31. Marking Accuracy	22	34	08	06	04	21	07	−12	−03	03	02	−01
32. Reading Comprehension G	05	17	53	61	47	07	19	27	43	06	26	34

[1] Decimal points omitted.
[2] For code numbers see table 10.10.

the Integration Battery (N = 266)[1]

13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
14	22	22	22	01	13	34	19	13	16	13	11	15	11	10	02	09	15	22	05
30	29	32	24	06	08	34	22	25	20	22	29	23	19	21	07	18	27	34	17
22	27	35	37	18	01	18	01	19	23	27	24	28	22	26	−08	31	18	08	53
25	29	45	41	22	15	15	28	22	33	44	42	31	31	34	−00	46	16	06	61
28	24	45	46	23	12	20	26	23	37	32	41	31	35	38	−03	44	14	04	47
22	20	22	11	−00	09	21	01	28	19	24	29	16	10	27	−04	26	41	21	07
28	17	28	14	00	15	23	10	38	27	34	34	21	14	20	03	35	41	07	10
−01	−01	01	15	22	10	−08	11	07	05	−00	−09	−03	−05	01	20	−03	−19	−12	27
09	24	24	40	22	18	12	31	02	25	10	03	13	07	15	18	12	−14	−03	43
27	29	30	20	12	11	29	18	21	22	19	22	20	14	18	33	24	16	03	09
25	20	37	41	03	11	32	33	29	29	20	32	25	30	31	12	32	07	02	26
36	36	46	43	07	14	25	36	40	42	32	55	38	38	41	07	47	35	−01	34
....	45	28	30	04	16	29	27	31	32	29	35	32	26	21	12	33	32	18	21
45	....	34	42	19	23	33	33	20	38	23	28	28	25	27	13	32	27	15	23
28	34	....	63	20	16	34	31	27	39	32	45	36	32	39	07	42	30	08	45
30	42	63	....	26	17	35	40	20	37	30	35	27	28	38	22	36	11	05	43
04	19	20	26	....	06	06	12	02	14	13	02	12	02	08	15	06	−07	−11	23
16	23	16	17	06	....	20	18	16	13	10	11	13	05	05	08	18	13	01	08
29	33	34	35	08	20	....	27	23	28	24	29	27	17	19	08	29	16	15	10
27	33	31	40	12	18	27	....	18	37	21	32	33	33	27	06	29	08	−01	26
31	20	27	20	02	16	23	18	....	33	39	40	21	23	21	00	32	28	05	18
32	38	30	37	14	13	26	37	33	....	33	35	36	37	39	04	40	11	−01	27
29	23	32	30	13	10	24	21	39	33	....	46	32	34	30	−01	42	20	07	33
35	28	45	35	02	11	29	32	40	35	46	....	34	37	34	03	45	30	09	30
32	28	36	27	12	13	27	33	21	36	32	34	....	34	52	−03	53	17	01	28
26	25	32	28	02	05	17	33	23	37	34	37	34	....	41	−05	48	10	01	31
21	27	39	35	08	08	19	27	21	39	30	34	52	41	....	−03	09	22	01	23
12	13	07	22	15	08	08	06	00	04	−01	03	−03	−05	−03	....	−06	03	−05	09
33	32	42	36	06	18	29	29	32	40	42	45	33	48	09	−06	....	33	07	33
32	22	20	11	−07	13	16	06	28	11	30	30	17	10	22	03	33	....	37	04
15	15	06	05	−11	01	15	−01	05	−01	07	09	01	01	01	−05	07	37	....	−05
21	23	45	43	23	06	10	26	15	27	33	30	28	31	23	00	35	04	−05	....

Table 10.10—*Centroid factor loadings for the Integration Battery*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	h^2
1. Speed of Identification, CP610A	31	-26	-29	-24	-19	16	-20	-08	-20	-03	07	-04	10	47
2. Spatial Orientation I, CP501B	47	-34	-16	-20	-23	15	-07	-18	-11	13	03	-13	07	57
3. General Information (TVN), CE505D	49	19	16	24	-30	10	11	-12	-06	15	-13	17	10	63
4. Reading Comprehension, CI614HX5	61	27	13	19	-15	11	01	-16	13	-15	-20	09	-10	66
5. Mathematics B, CI206C	62	24	20	20	-16	-19	-11	10	04	07	07	11	-07	63
6. Numerical Operations, Front, CI701B	36	-38	40	16	-19	-19	-07	21	-07	-03	12	-08	06	63
7. Numerical Operations, Back, CI701B	49	-31	36	30	07	-11	-16	09	-05	-05	19	-03	06	66
8. Mechanical Information, CI905A	08	33	-29	41	16	20	-10	15	08	21	02	-10	07	53
9. Mechanical Principles, CI903A	31	41	-43	21	10	13	-12	18	07	-06	-04	06	-05	58
10. SAM Complex Coordination, CM701A	41	-18	-19	10	21	-17	24	-15	-14	-03	12	13	-09	46
11. Planning Air Maneuvers, CI408AX3	49	16	-14	-13	18	-05	-07	-26	17	05	11	20	04	50
12. Planning A Course, CI406AX2	66	03	20	-12	11	-20	03	-07	18	08	04	-13	11	62
13. Instrument Comprehension I, CI615A	55	-21	-03	-06	09	-03	14	10	12	08	-18	08	09	46
14. Instrument Comprehension II, CI616B	57	-09	-21	-06	-04	-06	21	19	-03	-04	-23	06	11	54
15. Figure Analogies, CI212AX1	68	14	-08	-06	-17	-17	-05	-12	02	-09	12	-09	-12	61
16. Spatial Visualization I, CI204AX2	65	28	-25	-08	-22	-17	-11	02	12	-10	-05	-04	-08	69
17. Map Distance, CP626B	22	20	-18	26	-13	-07	11	04	-14	03	-09	-17	-09	29
18. Hands V-1	27	-08	-14	09	08	07	-04	13	07	-19	05	04	10	19
19. Cubes V-2	49	-21	-22	-12	-02	-07	-08	-08	-18	-17	08	16	11	47
20. Route Planning, CI411AX1	49	16	-21	-21	21	-07	-11	09	11	-14	-05	-07	06	47
21. Organizational Planning, CI407BX	49	-18	16	07	22	08	-20	-13	08	09	-05	-14	10	46
22. Following Oral Directions, CI651AX3	39	12	04	-04	12	-10	-04	07	-06	-06	-13	-05	07	44
23. Following Directions, CP602A	54	-04	27	06	05	05	-10	-14	-08	-12	-30	-08	-22	54
24. Code Analysis, CI653AX2	62	-06	24	-10	08	-09	-09	-22	18	-03	-08	-06	-12	58
25. Flight Formations, CI654AX5	55	15	14	-19	10	12	15	04	-15	-10	01	-05	06	47
26. Forced Landings, CI652AX4	49	18	18	-22	11	-01	-04	00	-14	21	-22	13	-21	53
27. Signal Interpretation, CI656AX2	58	30	24	-27	04	19	18	08	-14	-05	19	-06	-15	72
28. SAM Two-Hand Coordination, CM101A	12	-11	-34	22	27	-11	22	-08	-04	11	16	-07	-16	40
29. Combat Planes, CI655AX5	68	16	34	-16	11	19	14	09	-03	-13	16	14	-09	78
30. Log Book Accuracy X1	30	-41	22	-08	-07	10	13	14	24	07	14	-03	-10	54
31. Marking Accuracy X1	16	-36	-04	-15	-30	21	-07	14	13	13	03	11	-18	42

[1] Decimal points omitted. Variable 32 in table 10.9 was a test of reading comprehension very similar to that included as variable 4 and was apparently not included in the centroid analysis of the matrix.

TABLE 10.11.—*Rotated factor loadings and communalities for the Integration Battery*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	h^2
1. Speed of Identification[2]	09	66	02	−04	01	04	02	00	05	−03	10	04	07	47
2. Spatial Orientation I	−01	62	09	−04	17	−02	12	07	21	01	17	19	12	57
3. General Information (TVN)	03	10	10	02	71	03	03	−04	03	26	00	14	−02	62
4. Reading Comprehension	30	−03	12	04	60	16	26	00	07	19	22	−01	−02	65
5. Mathematics B	19	−02	38	08	40	04	00	02	04	50	16	−01	04	63
6. Numerical Operations, Front	−04	11	68	−18	12	01	02	−04	23	14	−09	11	08	62
7. Numerical Operations, Back	−05	08	75	04	17	07	08	07	10	07	04	−03	11	65
8. Mechanical Information	15	−09	−02	67	11	01	−09	−04	−11	02	01	09	−06	53
9. Mechanical Principles	47	02	−11	49	16	13	−05	−05	−10	15	09	−09	05	57
10. SAM Complex Coordination	12	11	15	03	15	06	02	45	00	00	05	−04	42	46
11. Planning Air Maneuvers	07	11	−01	13	24	01	06	13	−06	17	49	−19	27	49
12. Planning A Course	02	01	32	00	20	06	17	00	11	28	46	16	34	62
13. Instrument Comprehension I	10	13	15	07	21	−07	21	−06	24	10	05	02	48	45
14. Instrument Comprehension II	28	21	04	03	21	11	11	−07	14	16	−01	13	50	52
15. Figure Analogies	36	22	20	−05	27	08	04	17	07	33	38	11	11	61
16. Spatial Visualization I	53	20	05	04	26	01	03	−04	05	39	34	06	14	69
17. Map Distance	31	00	03	17	19	01	00	14	−06	14	−09	28	−01	29
18. Hands V-1	21	14	17	12	07	13	−02	−07	06	−09	06	−09	18	19
19. Cubes V-2	20	45	17	−10	14	06	00	10	−06	03	11	−11	31	44
20. Route Planning	32	12	04	13	−03	17	15	−10	−03	18	36	−02	31	46
21. Organizational Planning	−08	16	38	20	14	02	35	00	06	00	28	05	15	46
22. Following Oral Directions	20	08	21	07	18	18	25	−06	−07	27	18	09	28	42
23. Following Directions	17	09	25	−01	26	15	55	10	07	16	10	05	02	54
24. Code Analysis	08	06	29	−07	23	03	40	10	17	20	42	00	16	59
25. Flight Formations	04	11	09	−02	19	46	14	01	02	22	21	13	22	45
26. Forced Landings	−05	07	−02	03	18	16	38	07	02	53	11	−05	14	53
27. Signal Interpretation	01	06	05	−02	17	59	06	11	15	41	30	10	06	69
28. SAM Two-Hand Coordination	12	−02	[illegible]	29	−05	−08	−06	46	04	−10	00	08	26	41
29. Combat Planes	−01	02	22	−03	31	57	12	06	20	33	28	−10	17	76
30. Log Book Accuracy XI	−11	13	32	−06	09	06	06	02	60	03	09	02	19	56
31. Marking Accuracy XI	−03	35	02	−04	05	−07	01	−06	50	06	−07	−10	−04	41

[1] Decimal points omitted.
[2] For code numbers see table 10.10.

The correlation matrix is presented for the 32 variables in table 10.9, the centroid loadings and communalities in table 10.10, and the rotated loadings for the 13 factors in table 10.11. The sample was composed of 266 classified pilots. At the time these pilots were classified, no considerable selection was made except on the basis of mechanical-experience, visualization, spatial-relations, and psychomotor-coordination factors.

The Factors

For each factor the list of tests in order of descending loadings is given. Only loadings of 0.25 or greater are included.

Rotated factor I is defined by the following data:

Test No.	Test name	Loading
16	Spatial Visualization I	0.53
9	Mechanical Principles	.47
15	Figure Analogies	.36
20	Route Planning	.32
17	Map Distance	.31
4	Reading Comprehension	.30
14	Instrument Comprehension II	.28

This is the well established visualization factor, which is conspicuous by its absence from integration tests. One might have expected it to some degree in all of them, especially in the Flight Formations test. The latter is apparently susceptible of successful execution without the aid of visualization of the type conspicuous in Mechanical Principles and Spatial Visualization I and II.

Rotated factor II is defined by the following data.

Test No.	Test name	Loading
1	Speed of Identification	0.66
2	Spatial Orientation I	.62
19	Cubes	.45
31	Marking Accuracy	.35

This is clearly the perceptual-speed factor, which almost always has loadings above 0.60 in the two leading tests. The only feature of interest here is that the Marking Accuracy test is about 12 percent a matter of perceptual speed, whereas from its appearance it would seem to be a rather pure test of speed of simple motor movement. The perceptual component must be attributed to the necessity of locating positions for marking and to the visual control of accurate manipulations of the pencil. No integration tests are significantly loaded with this factor, even though they are usually speed tests and require attention to details.

Rotated factor III is defined by the following data:

Test No.	Test name	Loading
7	Numerical Operations (Back)	0.75
6	Numerical Operations (Front)	.68
21	Organizational Planning	.38
5	Mathematics B	.38
30	Log Book Accuracy	.32
12	Planning A Course	.32
24	Code Analysis	.29
23	Following Directions	.28

This, the numerical factor, has somewhat lower factor loadings in the classification tests than usual. This may indicate that other tests also have reduced loadings in it. Since pilots were not selected by any tests strongly weighted for the numerical factor, restriction of range can hardly be blamed for this state of affairs. The integration and planning tests in the list all involve the use of numbers in an elementary fashion so that some degree of saturation with this factor is understandable. It is interesting to see how analysis separates sharply between Marking Accuracy and Log Book Accuracy with respect to this factor, as should have been expected from the fact that in the latter, item numbers were involved.

Rotated factor IV has significant loadings in three tests:

Test No.	Test name	Loading
8	Mechanical Information	0.67
9	Mechanical Principles	.49
28	Two-Hand Coordination	.29

Here the use of classified pilots and their restriction on mechanical variance is quite evident. This is the well-verified mechanical-experience factor. None of the integration tests have significant loadings in it.

Rotated factor V is defined by the following data:

Test No.	Test name	Loading
3	General Information (Navigation Score)	0.71
4	Reading Comprehension	.60
5	Mathematics B	.40
29	Combat Planes	.30
15	Figure Analogies	.27
16	Spatial Visualization I	.26
23	Following Directions	.26

This is the verbal factor, which has no serious loadings in any integration tests except Combat Planes, and in this test it accounts for only 9 percent of the total variance. Vocabulary or verbal comprehension is thus of trifling importance in these integration tests.

Rotated factor VI is common to but three tests:

Test No.	Test name	Loading
27	Signal Interpretation	0.59
29	Combat Planes	.57
25	Flight Formations	.46

Here is decidedly a new factor, characteristic of three of the integration tests. None of the three is a pure measure of it. Flight Formations comes nearest to being a pure measure, since its secondary loadings are insignificant. Its communality, however, is low, and its variance in this factor is only 21 percent. If one were to attempt to develop a pure test of this factor, either this form would be cultivated or one would attempt to rid the other tests in this list from the intolerable secondary variances. The factor, for the present, may be called integration I until more is known regarding it.

The chief thing that these tests have in common is the requirement for the examinee to memorize and to retain a number of rules which must be followed in responding to the items. From this consideration the variable might be defined as a memory factor. It is possible that there is a factor having to do with the retention of verbal instructions and that it is common to these tests and the Memory for Tactical Plans test (see ch. 11). No correlations are available with which to test this suggestion.

Rotated factor VII is another one prominent in some of the integration tests:

Test No.	Test name	Loading
23	Following Directions	0.55
24	Code Analysis	.40
26	Forced Landings	.38
21	Organizational Planning	.35
4	Reading Comprehension	.25
22	Following Oral Directions	.25

The distinguishing feature of the leading tests in this list is apparently an ability to adapt quickly to new instructions and to carry them out successfully. Almost every item introduces new variations or modifications of general instructions given at the beginning. There is some necessity to retain mental sets, but not for nearly so long periods as in the case of factor VI just described. It would be highly desirable, however, to correlate these tests with memory tests in order to determine the possible identity of this factor with some memory factor. Until further information is forthcoming, it is best to name this factor integration II.

Rotated factor VIII is strong in only two tests:

Test No.	Test name	Loading
28	Two-Hand Coordination	0.46
10	Complex Coordination	.45

In spite of the very small number of tests with which to define this factor, it is probably the psychomotor coordination factor, held in common with the Rotary Pursuit Test and Finger Dexterity as shown in other analyses (see ch. 28).

Rotated factor IX is also restricted to two tests:

Test No.	Test name	Loading
30	Log Book Accuracy	0.60
31	Marking Accuracy	.50

This factor might be called marking speed, in accordance with one simple, obvious aspect of these two tests. It may be some independent type of clerical ability, but there is also the possibility that it is a much more restricted ability, such as speed of simple motor reactions as was found in certain factor-analysis studies preceding the war.

Giving credence to the last-mentioned hypothesis, the writer favors the name psychomotor speed for this factor.

Combat Planes, the highly speeded integration test which involves direct marking of the answer sheet, also has a small loading on this factor. It is clear that psychomotor speed does not enter into other integration tests to any appreciable degree and so does not add to the complexity of the tests, as had been feared.

Rotated factor X involves a number of integration tests:

Test No.	Test name	Loading
26	Forced Landings	0.53
5	Mathematics B	.50
27	Signal Interpretation	.41
16	Spatial Visualization I	.39
15	Figure Analogies	.33
29	Combat Planes	.33
12	Planning A Course	.28
22	Following Oral Directions	.27
3	Technical Vocabulary (navigation score)	.26

This is the usual general-reasoning factor. There are two other reasoning factors, each of a more restricted nature. This one, which has always been prominent in Mathematics B (arithmetic reasoning), shows up strongly in several of the integration tests. Forced Landings in particular appears to be as good a measure of it as Mathematics B. The items in forced landings are, after all, simple arithmetical-reasoning problems, in which the number work is so simple that the number variance drops out. As a result of the findings here, we conclude that a number of the integration tests would be valid for selection of navigators but would not be aided by reason of general-reasoning variance for the selection of pilots.

Rotated factor XI is defined by the following data:

Test No.	Test name	Loading
11	Planning Air Maneuvers	0.49
12	Planning A Course	.44
24	Code Analysis	.42
15	Figure Analogies	.37
20	Route Planning	.36
16	Spatial Visualization I	.34
27	Signal Interpretation	.30
21	Organizational Planning	.28
29	Combat Planes	.28

This factor list includes a combination of integration and reasoning tests. It seems to be identical with a factor isolated in two other analyses of the Nonverbal Reasoning battery (see ch. 7) and the Foresight and Planning II battery (see ch. 9). The tests, with two possible exceptions (Figure Analogies and Spatial Visualization I), seem to have in common the necessity for keeping in mind a number of detailed considerations provided either from the instructions or from the objects used in the items. Failure to take into account all considerations leads almost inevitably to an incorrect response. One hypothesis would be that it is a span of apprehension or a scope of apprehension. Another might be that it involves mastery of details. A third hypothesis, somewhat different from the other two, is that the factor is ideational fluency, the ease with which the individual can think of new possible responses. This ability would provide a distinct advantage in most of these tests except Code Analysis, Planning a Course, and perhaps Signal Interpretation and Combat Planes. Until further definitive evidence is available, it is best to name the factor integration III.

Rotated factor XII is defined by a single test, namely, Map Distance, which has a loading of only 0.28. This factor might have been regarded as a residual, except for the fact that there is too much spread in the loadings and it is possible to find concordant results in other analyses. This leads to the suggestion that it is the length-estimation factor in which Map Distance has previously shown a loading of 0.31. Spatial Orientation I is the only other test in the present battery that has a loading with the factor approaching significance. A small amount of length estimation in this test could be rationalized.

Rotated factor XIII is defined by the following data:

Test No.	Test name	Loading
14	Instrument Comprehension II	0.50
13	Instrument Comprehension I	.48
10	Complex Coordination	.42
12	Planning A Course	.34
19	Cubes	.31
20	Route Planning	.31
22	Following Oral Directions	.28
11	Planning Air Maneuvers	.27
28	Two-Hand Coordination	.26

This is the spatial-relations factor originally called the "intellectual component of the Complex Coordination test." For classification tests the loadings here are somewhat lower than the normal levels, due undoubtedly to the selection of the pilots. The loadings in general might, therefore, be higher in an unselected sample. However this might be, it appears that both forms of Instrument Comprehension are better tests of the factor than is Complex Coordination. The Planning a Course test, which in some respects was to duplicate the fundamental nature of the Complex Coordination test on paper, did not measure up to its model with respect to the measurement of spatial relations. It had a loading of 0.62

in this factor in the foresight-and-planning analysis (see ch. 9), however, so that this statement must be made with reservations. Its loading here suffers somewhat from the general restriction in common with other space tests. It is interesting to note in passing that the Hands test did not have an appreciable loading in this space factor. Results elsewhere (see p. 417) will show that it better represents another space factor.

Conclusions

Several important deductions and implications can be drawn from the results of this analysis.

In the first place, the chief feature of the Complex Coordination test, and one of the aspects that makes it valid, is not an integration ability, as had been hypothesized. Its chief variances which contribute validity for pilot selection, and for so many other kinds of predictions, are its spatial-relations and psychomotor-coordination factors, and to a small degree its variances in visualization and perceptual speed.

In spite of the fact that the hypothesis was wrong, it was fruitful in leading to test development in a new area and to some understanding of that new area. The three integration factors uncovered by this analysis become the starting points for new explorations in individual differences. One factor seems to represent the effective persistence of a complicated mental set which operates in rapid, complex, clerical-type work (integration I). A second factor seems to represent an adaptability of mental set; the trait of being able to modify sets on short notice (integration II). One might be tempted to call it flexibility of set (absence of perseveration), except for the fact that flexibility (or perseveration) tests have notoriously failed to intercorrelate to any substantial degree (see ch. 20 for an example of this). The third factor may represent some kind of span or scope of apprehension or attention; the ability to keep all elements in a set operating effectively (integration III). It will require further test development to examine all these hypotheses effectively.

As a group, integration tests proved to be nonvisualizing (at least visualization of the manipulatory type), nonperceptual (in the perceptual-speed sense), mostly nonnumerical, nonmechanical, and mostly nonverbal. Neither are they given to variances in the psychomotor-speed factor. The only better known factors with which many of them are secondarily involved are general reasoning and spatial relations (to which reference has already been made). The involvement with general reasoning often comes about in many a complex task, particularly when difficulties are encountered and responses are not obvious by way of visualization.

CHAPTER ELEVEN

Memory Tests[1]

MEMORY IN AVIATION

Memory in Aviation Training

Training, no matter of what type, implies learning and memory. It is therefore appropriate to expect that memory ability or memory abilities should be important in many phases of the education of air-crew trainees. When the operations of their training are observed and analyzed, this conviction is even greater. Ground-school training requires the aviation student to absorb large quantities of factual material under pressure in limited time. He is expected to remember, and to use later in flying, the information that is imparted to him in his classroom work. In learning to operate within his specialty in the air, he must also acquire a great number of skills that he did not possess before. These skills must be stamped in with repeated drill, and, if possible, over-learned to the point where he may perform automatically on occasion, resisting the effects of distraction or of stress.

Memory in Combat Operations

In combat it is expected and hoped that what the air-crew member learned in the way of factual information and in the way of motor skills will be sufficiently retained, and reinstated with sufficient facility for him to perform the necessary operations for which he spent many months of preparation. It is also true that training never ceases after flying personnel have passed beyond the stages designated as training. In other words, to maintain proficiency and to improve proficiency, the individual must acquire new information and skills and must perfect skills that he could not practice completely before.

In addition to the maintenance and improvement of proficiency in his job, the flying soldier goes through periods of briefing in which he is expected to note and to remember the important facts concerning the mission he is about to fly. He must remember his orders and specific features of the mission that are not carried with him in the form of written or pictorial material. He must be able to identify features of the landscape, if necessary, as well as friendly and enemy aircraft that may appear. Returning from his mission, he should be able to remember and to relate to interrogation personnel the important details that they wish to know. From the beginning of training through to the "pay-off" in

[1] Written by T/Sgt. Gerald H. Shirley, with contributions by Lt. C. H. Patterson and Lt. Eli A. Lipman.

combat, successful performance would seem to depend to a very large degree on the efficiency of the memory of each flying individual.

Job Analysis Data

It is not necessary to enumerate one by one the many things that must be remembered during air-crew training. Such a listing of specific acts of memory would be superfluous and the role of memory can be taken for granted without it. Objective data concerning the relative importance of remembering, however, should be examined before a full conclusion is reached regarding its place in air-crew performance. There are a variety of data that can be cited in connection with various phases of training and combat.

Memory Deficiencies in Training

In the data on 1,000 primary pilot elimination-board cases, it was reported that 24 percent of the cadets exhibited memory deficiencies sufficiently serious to be mentioned. Some of the typical notations are as follows: "Does not retain instruction," "requires repeated demonstrations," "repeats mistakes from day to day," "forgets fuel dial after one look," "fails to switch tanks," "forgets wind direction," "forgets flaps," "forgets to look back at tee on take-off," "forgets to notice tee on landing," and "neglects reference points on wing." In another sample of 1,303 primary eliminees for whom ratings had been given on the pilot rating scale of the Air Force Training Command, 39 percent were checked as having memory deficiencies.

No data are available for basic training, but in advanced single-engine training 52 out of 100 eliminees were reported as showing memory defects, and this represented 8 percent of all comments made concerning them. In advanced twin-engine training, 38 out of 100 had mentions of memory deficiencies, which represented 7 percent of all comments. In operational training, of 100 pilots who were reclassified for insufficient proficiency, 3 percent were reported as having memory defects.

Memory and the Bomber Crew

The relative importance of memory, as judged by supervisory officers of combat personnel in the Eighth Air Force, is indicated by ratings made on a scale of nine points. In evaluating the importance of memory for bomber pilots, the average rating, as judged by 74 observers, was 6.4 when the range of average ratings for other qualities was from 4.1 to 7.5. The average rating for navigators, as judged by 57 observers, was 6.9 when the means of other traits ranged from 5.0 to 8.0. The average rating for bombardiers, as judged by 31 observers, was 7.0 with a range for other traits of 5.3 and 8.0. Memory ranked ninth among 20 psychological requirements for the bomber pilot, being tied with the trait called "estimation of speed and distance." It was tied for eighth place along with arithmetic calculations and leadership in the case of navigators, and for sixth place along with division of attention and finger dexterity for bombardiers. In other words, its position among traits in general is regarded as higher than average.

RESEARCH ON MEMORY TESTS

A Systematic Plan

While a considerable amount of experimental work has been done in the field of memory, and the techniques of memory research are numerous and well known, not a great deal is known concerning individual differences with respect to memory performance. Previous factor-analysis studies have at least demonstrated the fact that memory ability is not a single trait in which people differ, but rather that there are a number of memory abilities. (1) It is not apparent when one considers the many learning activities and the many situations calling for remembering in connection with combat aviation just what specific memory factors are most important. It seemed desirable, therefore, to make a rather comprehensive and searching survey of this area—within the time permitted by the urgency of the military situation—in order to be sure that the important memory variables would be investigated. This implies a "shot gun" approach, but it was not by any means a completely blind approach. There was not sufficient time to explore all possible avenues, and there were restrictions in terms of the type of memory task that could be suited to the routine of classification testing. Within these limitations, a rather extensive plan of research was evolved.

The plan included factor analyses of two batteries of memory tests in order to determine what fundamental variables were important and which tests were most saturated in them. Since the time available for validation testing was limited at this period of research, only those tests with the highest factor loadings in memory abilities were to be validated. As it happened, the opportunity to validate most of them was later provided.

Features of Memory Tasks

If one recalls the various laboratory techniques, such as memory span, paired associates, serial learning, and the like, and if one also considers the various types of materials and the many methods for measuring the amount of retention, recall, and recognition, the lines of test possibilities tend to become clearer.

Types of material.—The favorite types of material are few. Verbal material is, perhaps, the most common. It may be either meaningful or nonsensical, and it may be presented in printed or in oral form. Pictorial material may be either schematic or photographic, and in meaningful or in relatively meaningless form. The things to be memorized may be presented in any of these forms in group testing, and the final test of efficiency of retention, recall, and recognition may also be given in terms

703320—47—16

of any of the same types of material. Presentation and final test activity may be in terms of the same kind of material and the same modality or type of perceptual or motor performance, or they may differ in these respects in many combinations. In selecting test ideas for development, the kinds of combinations most common to air crew were given highest priority.

Within the limitations imposed by group testing, it was not possible to set up the conditions for assessing individual differences in the retention and use of motor skills and other habits.[3] The approach described in this chapter voluntarily restricts itself to the type of memory task that can be applied in group testing. It will be found that the tests that follow depend frequently upon learning by paired associates. The material is either pictorial or verbal. The final measurement of memory proficiency is in terms of recognition tests—multiple-choice and matching types of response. This restriction was imposed by the use of the answer sheet. It is upheld by the conviction that efficiency of recognition and of recall are very highly correlated, at least within the limits of the retention intervals utilized in the tests.

Immediate vs. delayed recall.—Most memory tests heretofore have required recall or recognition only after relatively short time intervals. Practical considerations have usually demanded this type of test. The memory involved in air-crew performance, however, is of the delayed rather than the immediate type. It may be that the two are not very highly correlated. Since classification testing of aviation students had to be completed within a 2-day period, it would have been possible to insert an interval of 24 hours. This would not have been very convenient, since all group testing was confined to 1 day. The only test utilizing more than a few seconds delay between observation and recognition test was one that involved an interval of approximately 2 hours. During this interval, students were occupied with other tests. Had the interval been longer, even extending through the noon mess period, there would have been opportunity for extraneous factors to disturb the reliablity of the recognition test. The intervening of nonstandard activities between the impression and test of retention and recall has always been a disturbing feature of long-interval memory tests.

Face Validity

It was quite easy to apply the principle of face validity to memory tests. Pictures of planes and their names, landmarks as seen from the air paired with names, and aerial maps to be remembered by name or by visual features, provided a wealth of material. A set of orders for a mission presented orally provided a simulated briefing. Identification of ships, aircraft, and of landmarks as in pilotage, were represented in sev-

[3] Other studies with apparatus tests have examined the relation between learning facility in psychomotor tasks and graduation from pilot training. These results will be reported elsewhere (report No. 4).

eral of the tests. Retention of verbal instructions was represented in others. The pertinence of most tests should be apparent to any observer who is conversant with military-aviation requirements.

Classification of memory tests.—The tests included in this chapter are classified logically on the basis of (*a*) type of material, (*b*) nature of the task involved, and (*c*) manner of presentation and response. There are two main kinds of material: (1) pictorial and (2) symbolic. Each of these main categories is subdivided into (1) tests which require memory for complex wholes and relations of parts and (2) those which require memory for simple wholes (paired associates). Each of these divisions may be further broken down according to manner of presentation of and response to the items.

PICTORIAL MEMORY TESTS

Placed under this category are those tests that involve the ability to remember and to recognize material of a nonverbal, pictorial nature, including both complex wholes and relations of parts, and simple wholes (paired associates). Tests involving memory for complex wholes and relations include those presenting a pictorial stimulus and pictorial response, and those presenting a pictorial stimulus with a verbal response or question. Tests involving memory for simple wholes include those presenting a pictorial-verbal stimulus and a pictorial-verbal response, where the memory is primarily for the pictorial element.

Rationale

As pointed out earlier in this chapter, air-crew personnel must carry with them, mentally, certain information necessary for the success of the mission. Much of this material is nonverbal, nonsymbolic, or pictorial. Some of this pictorial material consists of recently acquired and complex information, such as maps of enemy territory with landmarks, targets, and other identifying features. Orientation to the terrain and objectives of the mission requires the recognition of identifying features as previously studied in maps of the territory. In addition, after the mission is accomplished, memory for the events of the trip in terms of the territory flown over is important in the accurate evaluation of results. The measurement of this type of memory is attempted in the Map-Memory tests which follow and the Memory for Landmarks test, which requires the remembering and recognition of single identifying features.

Memory for complex wholes and relations is also important in the routine, mechanical performance of air-crew duties under conditions where attention is diverted by other necessary activities or distractions. Thus, the pilot must know his instruments and their relative positions by memory, so that he can operate them without seeing them, when necessary. This type of memory is represented by the Memory for Instrument Board test described below.

Another type of pictorial memory is necessary for the quick recognition and differentiation of enemy and friendly aircraft and ships. Although to a certain extent proficiency in this area requires constant learning as new types of aircraft and ships are put into combat service, much of it represents the memory for outlines, forms, and identifying characters learned early in training. This area is sampled by the Memory for Planes and Memory for Ships tests described in this chapter.

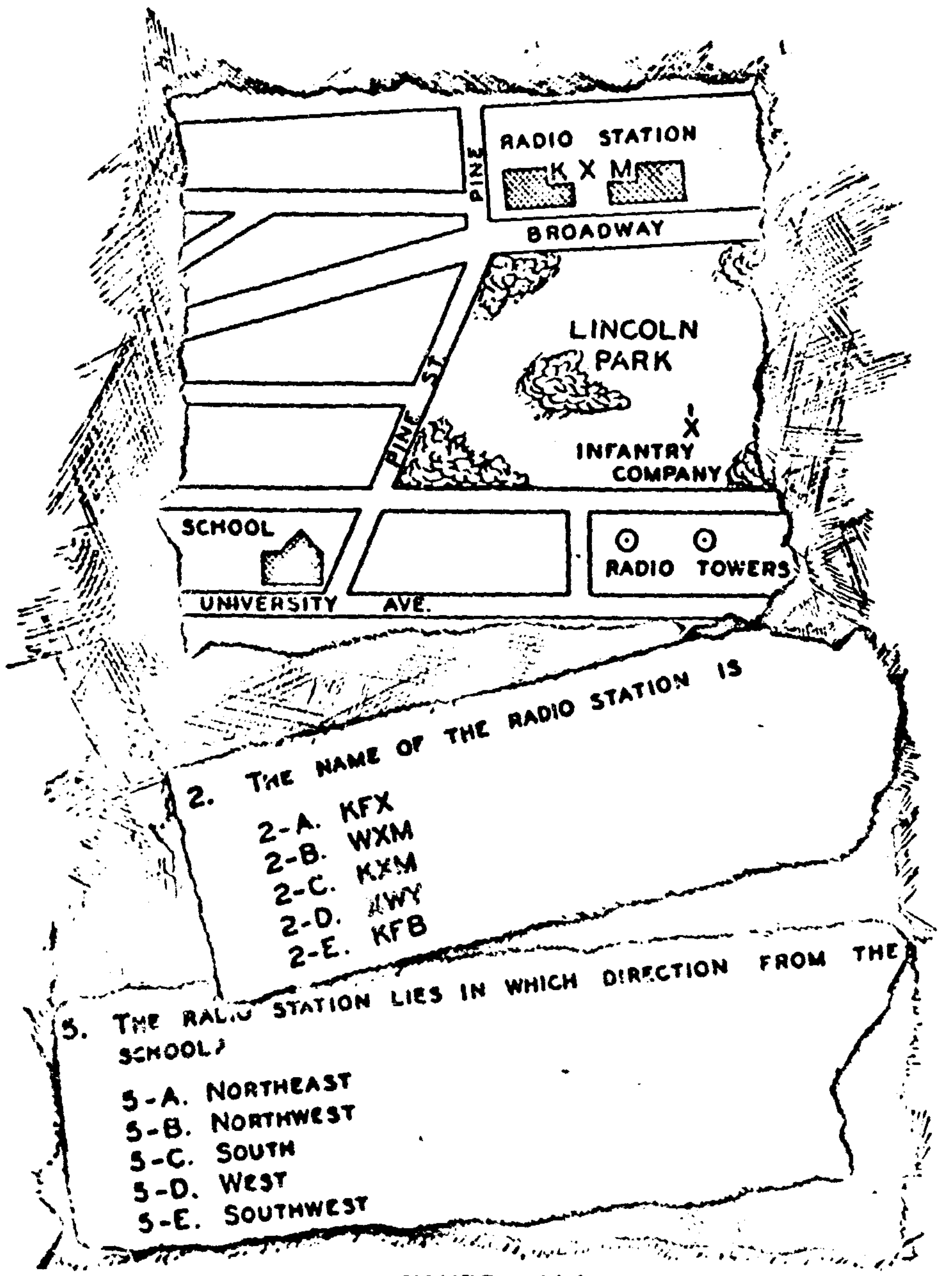


FIGURE 11.1
STUDY MAP AND TEST ITEMS OF MAP MEMORY, CI 505AX1

Map Memory, CI505AX1 [a]

This is a pictorial-memory test in which the stimulus is pictorial and the response is in verbal form. It was designed to measure the ability to remember complex wholes and relations of parts.

Description. (1) *Internal characteristics.*—The test consists of 2 parts, each in a separate booklet containing 60 items. Part I contains a sample map with 6 practice items and 3 test maps, each followed by 20 items concerning each map. Part II consists of 3 additional maps of the same type followed by 20 items each. Each item has five alternative verbal responses which distinguish this test from the visual form described later. Sample questions and a section of the appropriate map are shown in figure 11.1.

(2) *Administration.*—Instructions for the test inform the examinee that the test is a measure of his ability to remember details of a map which he is to study for a brief period of time. He is directed to note particularly, in his study of the large map, features which will enable him to remember any section, such as:

1. Names of places and things.
2. Locations of places and things in relation to each other.
3. Compass directions, e. g., location of one part of the map as north of another part.
4. Important routes by road, rail, air, etc., from one part of the map to another.
5. Number of times certain important objects occur in the map. The total time limit is 45 minutes for part I and 41 minutes for part II. Two minutes are allowed for study of the sample map in part I, followed by 2 minutes for answering the six sample items. Each of the test maps in part I is studied for 4 minutes, and 8 minutes are allowed for answering each set of 20 items. In part II, 5 minutes are allowed for the study of each map and 7 minutes for answering each group of 20 items. The administration of part I and part II of the test is separated by the administration of other tests in the battery in order to decrease the effect of proactive inhibition or other interferences.

(3) *Scoring.*—The scoring formula is $R-W/4$.

Statistical results.—The data given below are for a group of 179 classified bombardiers and 259 unclassified aviation students tested at Psychological Research Unit No. 3 in September 1942.

(1) *Distribution statistics.*—For 358 cases (bombardiers and unclassified students), the mean raw score (for parts I and II, 120 items) was 64.8, with a standard deviation of 18.2 and a range from 18 to 105.

(2) *Internal consistency.*—For the bombardier sample, the phi coefficients ranged from −0.10 to 0.55, with a mean of 0.33 and a standard

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Harry Rosenberg and Test Construction Staff.

deviation of 0.11 for part I. For part II, the range was from 0.01 to 0.57, with a mean of 0.30 and a standard deviation of 0.11. For an N of 179 (upper and lower halves of approximately 90 were employed) a phi of approximately 0.14 is required for significance at the 5 percent level, and a phi of 0.20 for significance at the 1 percent level. Although, as a whole, the test showed fair consistency, many items were lacking in discrimination between the upper and lower criterion groups.

(3) *Reliability coefficients.*—The correlation of part I with part II was 0.71, which yields a coefficient of 0.84 when corrected for double length. The group of 179 bombardiers was used for the computation.

(4) *Difficulty.*—Difficulty indices (N=179) ranged from 0.19 to 0.96, with a mean of 0.58, corrected for chance, and a standard deviation of 0.17 for part I; and from 0.21 to 0.93, with a mean of 0.61, corrected for chance, and a standard deviation of 0.16 for part II. Difficulty is therefore satisfactory, although some of the items are perhaps too difficult and others somewhat too easy.

(5) *Factorial composition.*—The most significant loadings are in the visual-memory (0.54), verbal (0.42), perceptual-speed (0.35), and visualization (0.26) factors. The communality is 0.70. For a fuller picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—A sample of 212 pilots yielded a biserial correlation of −0.16 between performance in this test and graduation-elimination in pilot training. The mean score for the graduates was 29.00, for eliminees 32.05, and the standard deviation for both combined was 11.04. Of this sample 75 percent were graduates.

Evaluation.—Factor analysis of this test shows that 70 percent of total variance is accounted for by common factors. The visual-memory factor accounts for 29 percent, the verbal factor for 18 percent, the perceptual-speed factor for 12 percent, and the visualization factor for 7 percent. The remaining 4 percent of the total variance is accounted for by common factors on which the loadings are quite low. Since the reliability is 0.84, the test contains some unknown common-factor variance.

The obtained validity of −0.16 is not congruent with that obtained for a similar form, CI505BX1 (see discussion immediately following). For the sample on which this validity was obtained, the standard error of the biserial correlation is 0.09. The −0.16, therefore, is not significantly different from zero.

Since the time required for the entire test (85 minutes) is impracticable for a single test, it is considered unnecessary to retain two parts. The test was considered worth revising because of the fact that certain items were highly related to total score. It was, therefore, purified by eliminating items with negative and low phis, as determined in the sample of 179 bombardiers. The resulting test is called Map Memory, CI505BX1, which is described next.

Map Memory, CI505BX1 [4]

This test is the first revision of Map Memory, CI505AX1. The only difference between the two tests is in length; length was reduced by eliminating items that did not discriminate between examinees with high and low total scores on the original form.

Description. (1) *Internal characteristics.*—The test consists of three large schematic maps, the first being a sample map. The sample map is followed by 5 sample items, and each of the two test maps is followed by 20 items. A sample map and item will be found in connection with the description of the original form of the test.

(2) Administration.—Thirty-five minutes are allowed for the entire test. The sample map is studied for 2 minutes, followed by 2 minutes for answering the five sample items. Each test map is studied for 4 minutes, and 8 minutes are allowed for answering each set of 20 items.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results. (1) *Distribution statistics.*—The test was administered at Psychological Research Unit No. 3 in September and November 1942, to 2,148 aviation students for validation. Distribution constants for each set of 20 questions and for the total test are presented in table 11.1.

TABLE 11.1.—*Distribution constants for Map Memory, CI505BX1, based upon a sample of classified pilots (N=793)*[1]

Part	M	SD
I	10.9	5.2
II	14.6	3.7
I and II	25.5	7.6

[1] In classes 44B and 44C.

(2) *Internal consistency.*—Thirty of the forty items are identical with those of part I of CI505AX1. The phis for these 30 items range from 0.24 to 0.51, with a mean of 0.38.

(3) *Reliability Coefficients.*—Map I items were correlated with map II items, yielding a coefficient of 0.66, corrected for length, for an N of 500 unclassified aviation students.

(4) *Difficulty.*—Difficulty values for the 30 items taken from CI505AX1 range from 0.37 to 0.92, with a mean of 0.61, corrected for chance.

(5) *Factorial composition.*—The most significant loadings are in the visual-memory (0.52), paired-associates memory (0.41), general-reasoning (0.23), and perceptual-speed (0.22) factors. The communality is 0.56. For a full picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—The test was validated against graduation-elimination in primary pilot school, map I and map II being validated sepa-

[4] Developed at Psychological Research Unit No. 3.

rately, as well as the total test. Validation results are presented in table 11.2.

TABLE 11.2.—*Validity data for Map Memory, CI505BX1, based on the criterion of graduation-elimination from elementary training*[1]

Part	M_g	M_e	SD_t	r_{bis}
Map I	10.92	9.82	5.18	.10
Map II	14.72	13.02	3.68	.22
Total	25.42	22.60	7.58	.17

[1] $N_t=793$; $p_g=.94$; classes 44B and 44C.

Evaluation.—By reason of the method of construction, i. e., selecting the best items from CI505AX1, the internal consistency of this test is relatively high. Its reliability, however, is only fair, perhaps because of its short length and/or its level of difficulty. Its total validity is not high, but it has much unique valid variance to offer, since both memory factors have some small validity for the pilot.

Factor analysis of this test shows that 56 percent of the total variance is accounted for by common factors, leaving a small amount of unknown variance. Of the known variance, 5 percent is accounted for by the perceptual-speed factor, 5 percent by the general-reasoning factor, 27 percent by the visual-memory factor, and 17 percent by the paired-associates memory factor. The remaining 2 percent is accounted for by factors on which the loadings are quite low. It is not a pure test, but its two leading factors both seem to be weighted in the pilot criterion.

An estimated validity coefficient (computed from factor validities; see table 28.17) is lower than that found empirically. This indicates that there is common-factor variance, valid for pilot training, that was unaccounted for in the analysis. This is probably visualization, which did not emerge in the battery in which the BX1 form was analyzed, but did emerge in the other memory battery, in which the AX1 form appeared.

Map Memory, CI505AX3 [a]

This test is similar to the two preceding tests and was designed to measure the same functions. It differs from Map Memory, CI505AX1 and CI505BX1, in that the period of study of the maps is reduced and fewer questions are asked on each map.

Description. (1) *Internal characteristics.*—The test consists of one sample map and six test maps of the same type used in the preceding tests. Four items follow the sample map, and 10 items follow each of the 6 test maps.

(2) *Administration.*—The total time for the test is 35 minutes. Thirty seconds are given for the study of the maps, including the sample map; 2 minutes are allowed for answering the 4 sample questions, 4 minutes for the first 10 items and 3 minutes for each of the 5 remaining sets of 10 items.

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Milton Burdman, Capt. Harry Rosenberg, Lt. Mahlon B. Smith.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data given below are for examinees tested at Psychological Research Unit No. 3 in November 1942.

(1) *Distribution statistics.*—A sample of 240 unclassified aviation students and 179 bombardiers yielded a mean score of 23.2 and a standard deviation of 8.6. The distribution curves were approximately normal.

(2) *Internal consistency.*—Phi coefficients ranged from 0.09 to 0.50, with a mean of 0.30 and a standard deviation of 0.10. Upper and lower groups of 100 each (highest 27 percent and lowest 27 percent of the total) were used. For an N of 200, phis of 0.14 and 0.18 are required for significance at the 5 percent and 1 percent levels respectively.

(3) *Reliability coefficients.*—For a group of the 239 students the sum of the scores in the first, third, and fifth groups of 10 items was correlated with the sum of the scores in the second, fourth, and sixth groups of 10 items, yielding a coefficient of 0.67 corrected for length.

(4) *Difficulty.*—Difficulty indices were computed for the same sample used to determine internal consistency. They ranged from 0.21 to 0.92, with a mean of 0.54, corrected for chance success, and a standard deviation of 0.16.

(5) *Factorial composition.*—The most significant loadings are in the visual-memory (0.55), verbal (0.31), visualization (0.31) and spatial-relations (0.21) factors. The communality is 0.61. For a fuller picture of the factorial composition of this test, see appendix B.

(5) *Test validity.*—Validation data are shown in table 11.3.

TABLE 11.3.—*Validity data for Map Memory, CI505AX3, for a sample of pilots in primary training, graduation-elimination criterion (N_t=176, p_s=.66)*

Part	M_s	M_f	SD_t	r_{bis}
I[1]	11.78	10.44	5.02	.16
II[2]	12.72	12.10	4.96	.08

[1] First three maps.
[2] Last three maps.

Evaluation.—This test has fair internal consistency, and its item difficulties are on the whole satisfactory. Its reliability is somewhat low, being the same as that of CI505BX1 which had fewer items although requiring the same amount of time. The pilot validity of the test is low, but it makes a unique contribution.

Factor analysis of this test shows that common factors account for 61 percent of the total variance, leaving only 6 percent of the nonerror variance unknown. Of this, the verbal and visualization factors account for 10 percent each, the spatial-relations factor for 4 percent, and the visual-memory factor for 30 percent. The remaining 7 percent of the total variance accounted for is found in factors on which the loadings are quite low.

Estimation of the pilot validity of CI505AX3, by means of factor equations (see table 28.18), yields a coefficient similar to that found empirically. This indicates that all the valid factors of this test have been accounted for by the analysis.

Map Memory (Visual Form), CI505AX2 *

This test involves pictorial memory for complex wholes and for relations of parts with both stimulus material and response material being pictorial. It was designed to measure visual memory for map details. The

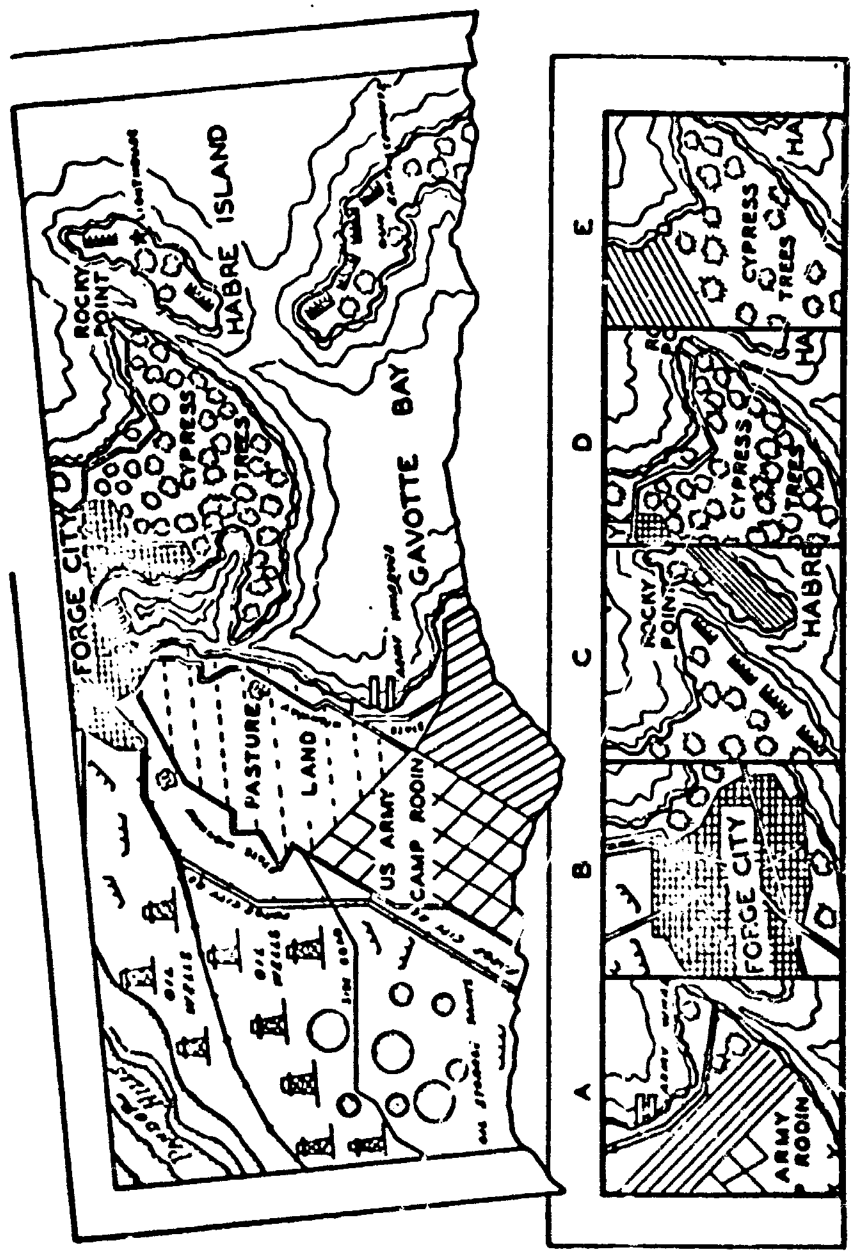


FIGURE 11.2
STUDY MAP AND TEST ITEM OF MAP MEMORY (VISUAL FORM), CI505AX2

* Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Milton Burdman, Sgt. Warren J. Cowan.

utilization of the recognition of pictorial response material rather than verbal questions makes the task more like that actually required of air-crew members in flight.

Description. (1) *Internal characteristics.*—The test consists of 3 large diagrammatic maps, each followed by 20 items, each item being in the form of 5 small map sections, one of which is an accurate reproduction of a section of the large map. A sample problem, consisting of a large map followed by an item, is shown in figure 11.2.

(2) *Administration.*—The instructions direct the examinees, when studying the large map, to note particularly features which will enable them to identify any section, such as:

1. Names of places and things.
2. Locations of places and things in relation to each other.
3. Number of times important objects appear in a given area.
4. Courses followed by roads, coastlines, boundary lines, etc.

The total time limit for the test is 60 minutes. Three minutes are allowed for study of a sample map, followed by 3 minutes for answering the three sample items. Five minutes are allowed for study of each of the 3 test maps, followed by 12 minutes for each set of 20 items.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—Data are available for pilots in class 44D tested at Psychological Research Unit No. 3 in November 1942; and for samples of navigators tested at Selman Field on May 31 and June 1, 1944, at Ellington Field on May 22 and 23, 1944, and at Psychological Research Unit No. 3 on May 4, 5, and 6, 1944.

(1) *Distribution statistics.*—The administration of the test to 689 classified pilots (tested while unclassified) yielded a mean of 29.2 and a standard deviation of 11.8.

(2) *Internal consistency.*—The phi coefficients ranged from 0.11 to 0.65, with a mean of 0.42 and a standard deviation of 0.10. For an N of 240 (highest and lowest fourths of 120 pilots each were employed) a phi of 0.13 is required for significance at the 5 percent level and one of 0.17 for significance at the 1 percent level. The results indicate satisfactory internal consistency for the test.

(3) *Reliability coefficients.*—The three groups of 20 items each were treated as separate parts and intercorrelated. The following reliability coefficients (corrected for triple length) were obtained (N=487 pilots): Part I vs. part II, 0.83; part I vs. part III, 0.79; part II vs. part III, 0.81.

(4) *Difficulty.*—Difficulty indices, computed for the same sample, ranged from 0.29 to 0.85, with a mean of 0.59, corrected for chance success, and a standard deviation of 0.13, indicating a satisfactory difficulty level for the sample studied (N=487).

(5) *Factorial composition.*—The most significant loadings are in the visual-memory (0.58), the perceptual-speed (0.35), and the verbal (0.23)

factors. The communality is 0.59. For a fuller picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Validation data are shown in table 11.4.

TABLE 11.4.—*Validity data for Map Memory, C1505AX2, based on the graduation-elimination criterion*

Group	Score	N_t	p_s	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$
Pilots in primary training . .	R−W/4	689	0.91	29.35	27.50	11.83	0.08	[1]0.11
Pilots through basic training	R−W/4	625	.86	28.90	28.14	11.66	.04	[2].10
Navigation students	Rights[3]	1,577	.91	39.31	35.57	9.75	.20	[2].34
Navigation students	Wrongs[3]	1,577	.91	15.48	18.53	8.45	−.18	[2]−.31
Navigation students	R−W/4	1,577	.91	35.69	30.94	11.35	.20	[2].34

[1] Assuming an unrestricted stanine standard deviation of 1.83.
[2] Assuming an unrestricted stanine standard deviation of 2.00.
[3] For this sample, the correlation between rights and wrongs is −.71.

(7) *Item validity.*—Based upon 671 pilots, 600 of whom graduated from primary training, item-validity phi coefficients ranged from −0.15 to 0.20 with a mean of 0.02 and a standard deviation of 0.04.

For an N of 671, phis of approximately 0.08 and 0.10 are required for significance at the 5 percent and 1 percent levels respectively.

Evaluation.—The estimated reliability of this test is satisfactory, and its correlations with other tests and with the pilot stanine are relatively low. It would contribute a very small unique pilot validity by virtue of its loading with the visual-memory factor, but at too great a cost in testing time. The average pilot validity is 0.17 (including data from one sample of 92 not mentioned before), which is almost all accounted for by known factors. The navigator validity coefficient is fairly high, which suggests that further exploration of memory tests for this air-crew position, particilarly those saturated with the visual-memory factor, would be worth while.

Factor analysis of this test has accounted for 59 percent of its total variance (compared with a reliability of about 0.80). Of this the verbal factor accounts for 5 percent, the perceptual-speed factor for 12 percent, and the visual-memory factor for 34 percent of the total variance. The remaining 8 percent is accounted for by factors on which the loadings are quite low. Map Memory, CI505AX2, is the purest test in this series, measuring the visual-memory factor fairly well. As such, it has value in factor-analysis research.

Visual Memory, CI514A [9]

This is a nonverbal memory test. It was designed to measure visual memory for parts of a complex whole. It was believed that the test would stress visual-memory ability to a greater degree than most forms of Map Memory, CI505. As in Map Memory, CI505AX2, response is made to a pictorial stimulus rather than to a verbal stimulus.

[9] Developed at Psychological Research Unit No. 3. Chief contributors: S/Sgt. Arthur Z. Cerf, Sgt. Hyman Heller, Pfc. Charles W. Nelson.

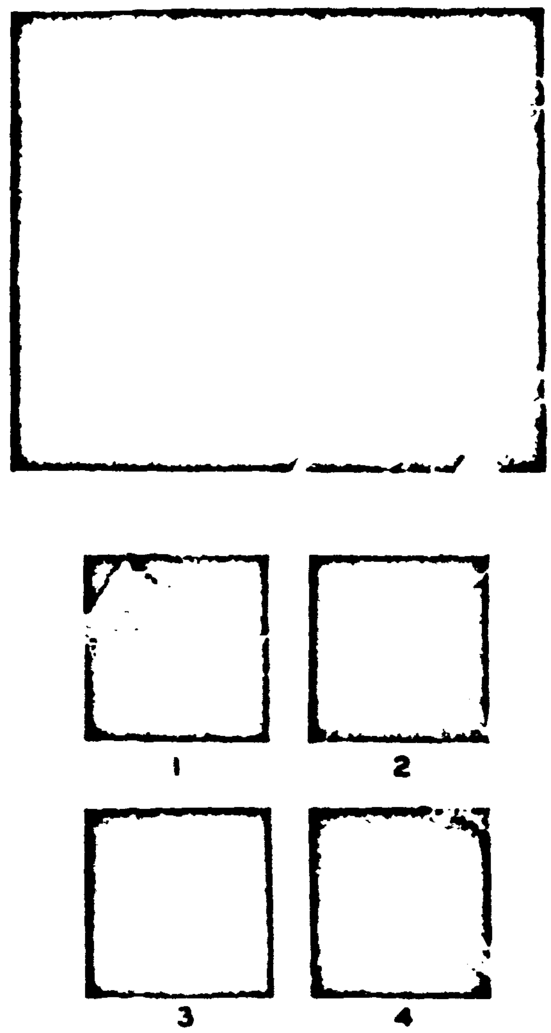


FIGURE 11.3
SAMPLE PHOTOGRAPH & FOUR TEST ITEMS OF VISUAL MEMORY, CI514A

Description. (1) *Internal characteristics.*—The test consists of five, page-size, aerial photographs (study-photographs), each with 24 small test-photographs. The examinee studies the large photograph, turns the page, then indicates which of the 24 small photographs are portions of the large one and which are not. A sample problem, consisting of a study-photograph and four test-photographs, is shown in figure 11.3.

(2) *Administration.*—The examinee is informed that the test is a measure of his ability to remember aerial photographs. One minute is allowed for studying each large map and 2 minutes for answering the 24 items. Examinees are told to follow their "hunches" in answering the items.

(3 *Scoring.*—The scoring formula is R−W+20.

Statistical results.—Data are available for unclassified aviation students tested in April 1945 at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—A sample of 298 unclassified aviation students yielded a mean score of 64.2 and a standard deviation of 14.8. The distribution curve is approximately symmetrical and somewhat more peaked than normal.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.44, a standard deviation of the phi distribution of 0.28, and a range of values from 0.00 to 0.98. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 750 unclassified aviation students.

(3) *Reliability coefficient.*—Preliminary evidence on reliability was obtained by the Kuder-Richardson method. An estimated reliability coefficient of 0.87 was obtained. This figure is based on a sample of 624 unclassified aviation students.

Evaluation.—The similarity between Visual Memory and the Map Memory forms, both in subject matter and presentation, leads to the belief that the test probably has approximately the same validity as those forms. This test is well constructed, and it would seem probable that it will prove a satisfactory instrument for measuring a type of visual memory.

Plane Position Memory, CI512A [a]

This test is also a nonverbal, visual-memory test. The test was designed to measure ability to remember parts of a complex whole, stressing memory for positions of objects.

Description. (1) *Internal characteristics.*—On each of four study-pages of the test are presented nine airplanes in three rows of three. The airplanes are of different types and are headed in one of four different directions (up, down, left, right) but all are shown from a side view. Following the study page, the nine airplanes are shown in different posi-

[a] Developed at Psychological Research Unit No. 3. Chief contributor: Cpl. Albert A. Canfield, Jr.

tions on another page and all are headed toward the left. Figure 11.4 shows a part of a study-page (in the upper panel) and the succeeding response-page (in the lower panel). There is one practice problem at the beginning of the test.

(2) *Administration.*—The examinees are informed that the test is a measure of their ability to remember the positions of planes, and that the task is to remember in what row the airplanes appear and in what direction they are headed. The examinees are given 2 minutes to study the planes; then, at a signal, the examinee turns to the response-page. The examinee is allowed 3 minutes to indicate, by marking A, B, or C, in what row the airplane appeared, and, by marking A, B, C, or D in the next-numbered space on the answer sheet, in which direction it was going. A box at the top of each of the response pages indicating, by means of arrows, the symbol for each direction, facilitates the examinee

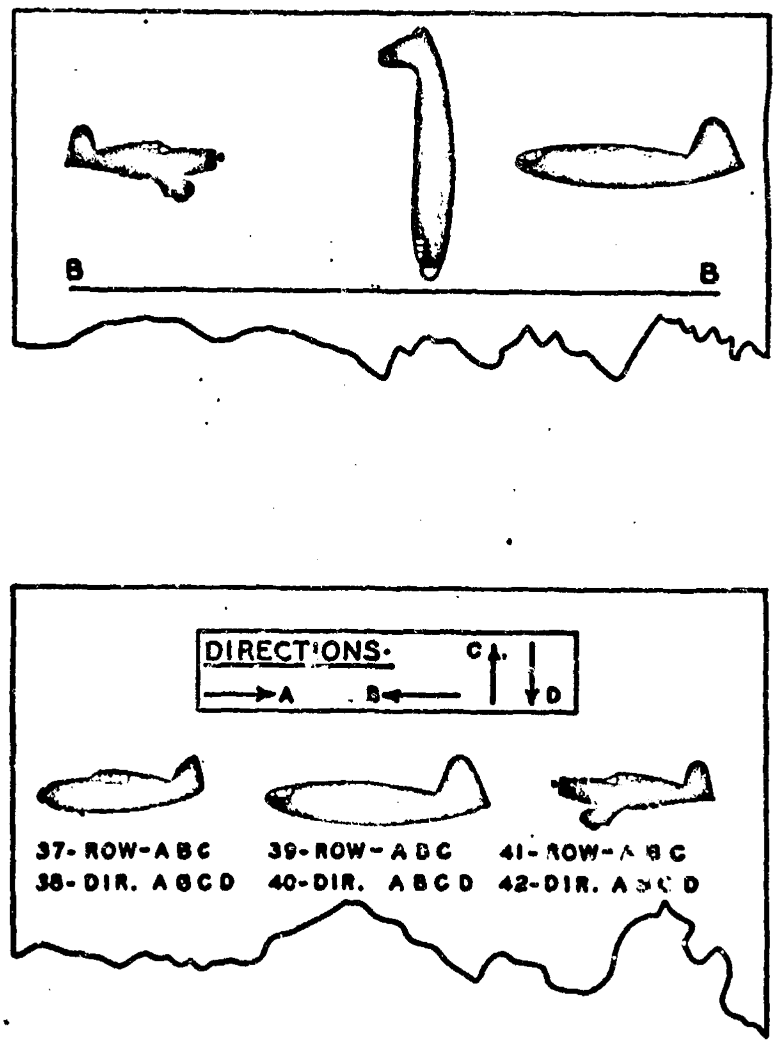


FIGURE 11.4
STUDY PAGE & SUCCEEDING RESPONSE PAGE OF PLANE POSITION MEMORY, CI512A

in marking answers. The total time for the test is 30 minutes with an actual testing time of 20 minutes.

(3) *Scoring.*—The scoring formula is R—W/3.

Statistical results.—No statistical data are yet available for this test.

Evaluation.—This test was designed to measure visual memory as purely as possible. Subjective examination of the test indicates that it should be fairly good for this function. It is a well-designed and executed test that, for use on an aviation-student population, has face validity.

Airplane Formation Memory, CI513A *

This is another nonverbal, visual-memory test, developed for analytical purposes. The test differs from Visual Memory, CI514A, not only in

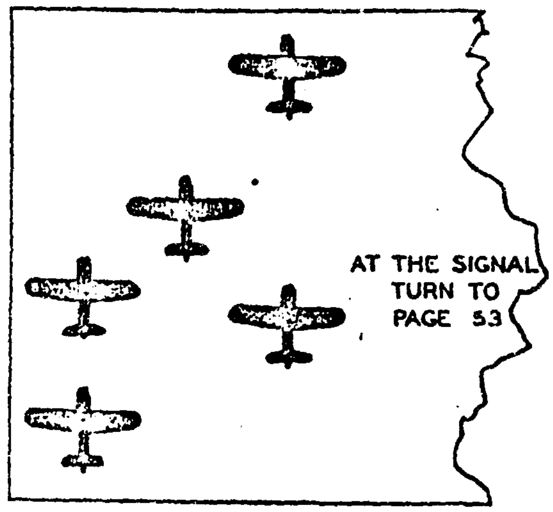
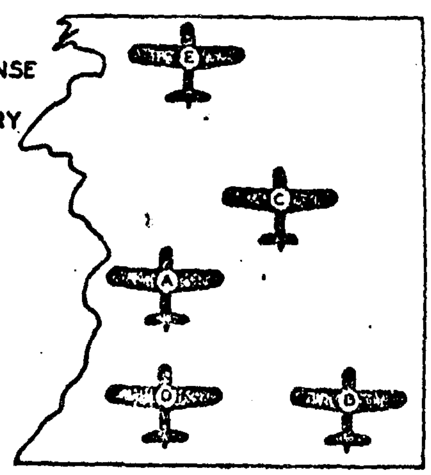


FIGURE 11.5
STUDY PAGE AND
SUCCEEDING RESPONSE
PAGE OF AIRPLANE
FORMATION MEMORY
CI513A



* Developed at Psychological Research Unit No. 2. Chief contributors: Cpl. Leland D. Brokaw, Florence R. Grossman, Lt. John I. Lacey.

subject content, but in that it measures ability to remember complex wholes.

Description. (1) *Internal characteristics.*—The test consists of two parts of 20 items each, with two sample items. The stimuli are diagrammatic planes in formation. The shape of the formation and the number of airplanes in it (4 to 10) varies from item to item. The shape of the planes and the view (top or side view) are constant within each item, though they vary among items. The response is made to another formation of planes similar to the stimulus-formation except that some planes have been moved out of position within it. The task of the examinee is to select those planes that have been moved out of position. A sample item, consisting of a stimulus-formation and the formation from which the response is made, is given in figure 11.5.

The examinees are given a brief time to study the formation (8 seconds) and then are told to turn to a given page in the back of test booklet where the response-formation appears. The response items are scattered randomly throughout the last half of the booklet. This was done in order to reduce the possibility of the examinee's getting answers by seeing the final positions of the succeeding problems, since there were two response-formations to a page. Thus, a few seconds elapse between the time the examinee has seen the original formation and has located the response-formation.

(2) *Administration.*—The examinees are told that this is a test of their ability to remember positions of airplanes in formation. After 8 seconds study for each formation, they are given the page and number of the response-formation (which is also printed on the study page), and allowed 15 seconds to locate, select, and blacken answers on the answer sheet. The total testing time is approximately 22 minutes, including the administration of the directions which take approximately 5 minutes.

(3) *Scoring*—The scoring formula is R—W.

Statistical results.—There are no statistics available for this test.

Evaluation.—Since there are no statistical data available for this test, it is difficult to make any real evaluation of it. The test was developed as part of the plan to construct pure factor tests. It is not known whether the visual-memory factor involves correct recall or recognition of details, or positions, or of forms or any combination of these. Hence a variety of such tests were developed for analytical study and for validation.

Memory for Plane Silhouettes, CI503AX1 [10]

This test is a nonverbal, paired-associates, immediate-memory test with a recognition response made to a pictorial stimulus. It was designed to measure ability to remember the relationships between paired wholes.

Description. (1) *Internal characteristics.*—Silhouettes of planes are presented on a page which the examinees study, the top and side-view

[10] Developed at Psychological Research Unit No. 3. Chief contributor: Lt. Frank J. Dudek.

silhouettes of each plane being shown paired. After a brief study period, the top silhouettes are presented in a column on a test page. In a parallel column, side-view silhouettes of the planes are shown on this page, but not paired with the top silhouettes as on the study page. A brief matching test is thus presented. More side-view silhouettes are presented in the recognition group than there are pairs on the study page. This is done in order to lower the dependence of one item on another, and thus prevent some right responses merely by a process of elimination. The task of the examinee is to pair up the silhouettes in the same way they were presented on the study page. There are 28 items involving 5 study pages, each with 4 to 8 pairs. Since the planes are different in each group, each section is independent of the others. There is a sample item at the beginning of the test. A sample item, consisting of two sets of planes, and three side-view silhouettes to which the responses are made, is presented in figure 11.6.

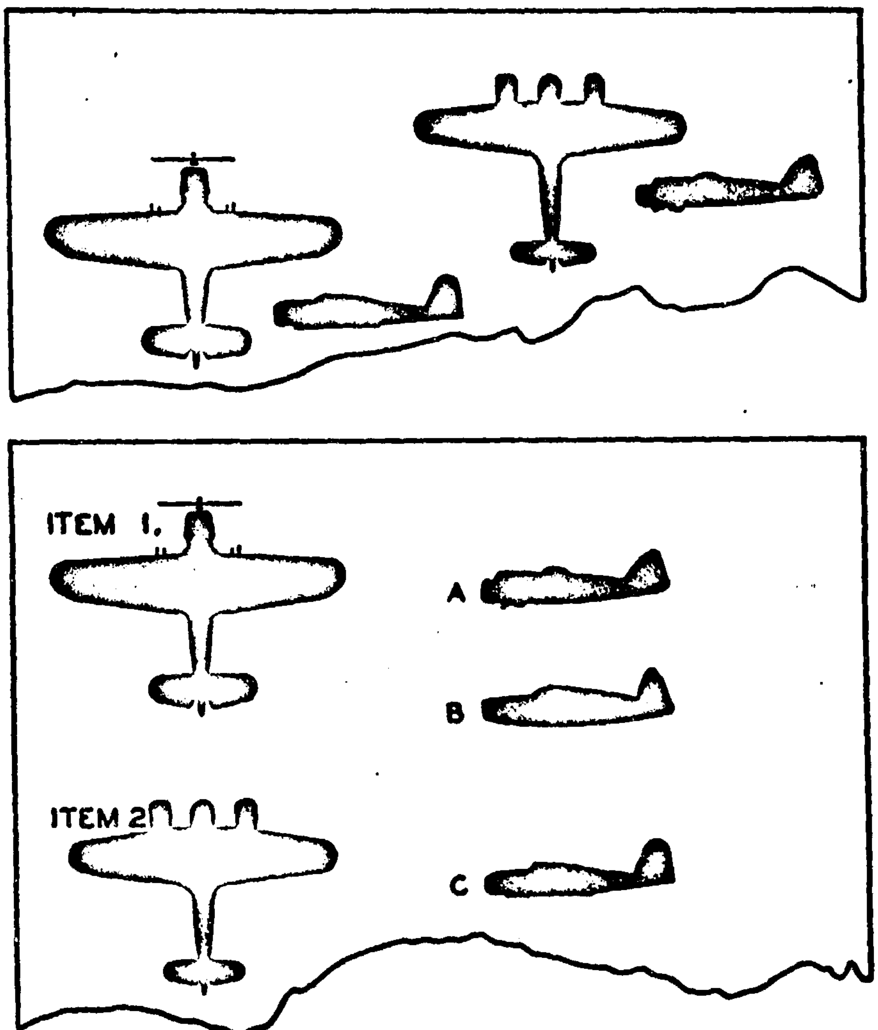


FIGURE 11.6
PORTION OF STUDY PAGE & TEST ITEMS OF MEMORY FOR PLANE SILHOUETTES, CI503AX1

703320—47—17

(2) *Administration.*—Eighty seconds are given for each study period. At the end of that period, a signal is given for the examinee to turn to the response page. Two minutes are allowed to match the planes and mark the answers. At the end of that period, another signal is given to turn to the succeeding study page.

(3) *Scoring.*—The scoring formula is R−W/5.

Statistical results.—Except where noted below, the following data are for unclassified aviation students tested in October 1942 at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—A sample of 238 unclassified aviation students yielded a mean score of 21.7 and a standard deviation of 5.0. The distribution curve is approximately symmetrical and somewhat flatter than normal.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.48, a standard deviation of the phi distribution of 0.10, and a range of values from 0.24 to 0.66. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 750 unclassified aviation students.

(3) *Reliability coefficient.*—By the alternate-forms method (part I-part II), an estimated reliability coefficient of 0.82, corrected for length, was obtained. This figure is based on a sample of 238 unclassified aviation students.

(4) *Difficulty.*—Based upon item analysis of the responses of 750 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.68, corrected for chance, with a range from 0.33 to 0.90 and a standard deviation of 0.14.

(5) *Factorial composition.*—The most significant loadings are in paired-associates memory (0.56), spatial-relations (0.38), and perceptual-speed (0.34) factors. The communality is 0.61. For a fuller picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Validation results based on several samples given are in table 11.5.

TABLE 11.5.—*Validity data for Memory for Plane Silhouettes, CI503AX1, based upon samples of pilots in primary training, graduation-elimination criterion*

N_t	p_s	M_s	M_e	SD_t	r_{bis}
[1]471	0.82	19.96	18.14	5.52	0.19
[2]169	.86	22.64	19.76	5.04	.31
[3]233	.70	19.64	17.90	5.76	.18

[1] Tested in January 1943 at Psychological Research Unit No. 3.
[2] Tested in November 1942 at Psychological Research Unit No. 3.
[3] Tested in October 1942 at Psychological Research Unit No. 3.

Evaluation.—Because this test has a fairly high correlation with the pilot stanine, it would apparently add very little to the classification battery in prediction of pilot success. This is due to the fact that the test has some loadings in perceptual speed (average for two analyses is

0.34), and spatial relations (average is 0.38), both being valid factors for pilot prediction and already heavily weighted in the pilot stanine. Its loading with paired-associates memory is a unique contribution, but it is not very heavily weighted in the pilot criterion.

The pilot validity of this test (weighted average of 0.21) is almost fully accounted for by its common factors. This test would be suitable to measure the paired-associates memory factor, except for the fact that it has significant foreign variance in the perceptual and spatial-relations

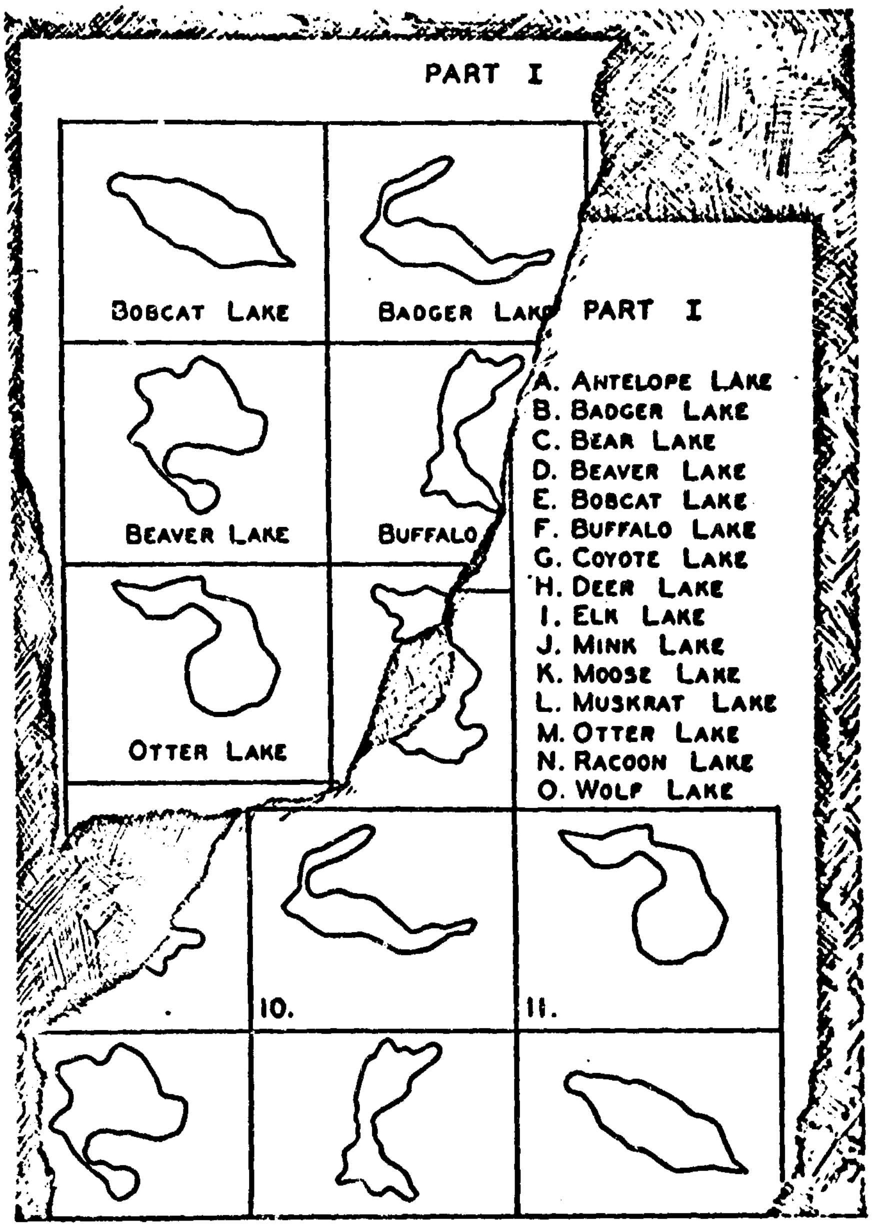


FIGURE 11.7
PORTION OF A STUDY PAGE AND THE CORRESPONDING RESPONSE PAGE OF MEMORY FOR LANDMARKS, CI510AX1

factors. Memory for Landmarks is a more pure and more heavily weighted test of the paired-associates memory ability and so is more suitable to represent it.

The known factor content of this test accounts for only 61 percent of its variance, compared to a reliability of 0.82.

Memory for Landmarks, CI510AX1 [11]

This is a visual-memory, paired-associates test in which a pictorial symbol is paired with a verbal symbol. In the recognition test following the study of these pairs, a long-matching-form arrangement is used.

Description. (1) *Internal characteristics.*—The test consists of 2 parts of 20 items each. Each part is in a separate booklet and is divided into two sections. For each section there is a study page on which 15 diagrams of similar topographical features, i. e., lakes, rivers, bays, are paired with their names. All the diagrams on any one page are of the same type of geographical feature, but the feature varies from page to page. After a brief study period, the examinee turns the page to the response material. This consists of 10 diagrams, identical with 10 of the 15 on the study page. Alongside these diagrams is presented the original list of 15 names, in mixed order. The task of the examinee is to match the names with the diagrams. There is a short practice problem at the beginning of the test. A sample of the diagrams on the study-pages and response-pages is given in figure 11.7.

(2) *Administration.*—The examinees are informed that this is a test to measure their ability to remember geographical landmarks. Four minutes are allowed for the study of each set of landmarks; then, at a signal, the examinees are told to turn the page and match the names with the landmarks. Three minutes are allowed for selection and marking of the answers. Strong cautionary statements are made prohibiting examinees turning back to the study page after the study period. The total testing time for each booklet is approximately 20 minutes.

(3) *Scoring.*—The scoring formula is the number right only.

Statistical results.—The data given below are for examinees tested in October 1942 at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 11.6. The distribution curves are approximately symmetrical and somewhat more peaked than normal.

TABLE 11.6.—*Distribution constants for Memory for Landmarks, CI510AX, based upon samples of unclassified aviation students*

Part	N	M	SD
Part I	366	7.1	3.8
Part II	366	8.9	3.9
Total	365	16.0	8.9
Total	233	16.0	8.9

[11] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, Lt. David H. Jenkins.

(2) *Internal consistency.*—Analysis of responses of several sample groups yielded the internal-consistency data given in table 11.7.

TABLE 11.7.—*Internal-consistency data for Memory for Landmarks, CI510AX based upon samples of unclassified aviation students*

Part	N	M_{phi}	SD_{phi}	Range of phi	
				Low	High
I	179	0.49	0.08	0.31	0.62
II	136	.40	.12	.25	.60

(3) *Reliability coefficient.*—By the alternate-forms method (part I-part II), an estimated reliability coefficient of 0.82, corrected for length, was obtained. This figure is based on a sample of 238 unclassified aviation students.

(4) *Difficulty.*—Based upon item analysis of the responses of 179 unclassified aviation students, indices of difficulty were found as shown in table 11.8.

TABLE 11.8.—*Difficulty indices for Memory for Landmarks, CI510AX, based upon 179 unclassified aviation students*

Part	M_{cp}	SD_{cp}	Range	
			Low	High
I	0.42	0.14	0.16	0.75
II	.50	.16	.24	.75

(5) *Factorial composition.*—The most significant loadings are in the paired-associates-memory (0.61) and visual-memory (0.20) factors, and in a third memory factor (0.44), which seems to be confined to this test and Plane Name Memory, CI506AX1. The communality is 0.68. For a fuller description of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Validation results are given in table 11.9.

TABLE 11.9.—*Validity data for Memory for Landmarks, CI510AX1, based on pilots in primary training (N_t=245; p_s=.62)*

Part	M_p	M_f	SD_t	r_{bis}
First section	3.62	3.62	2.23	0.00
Second section	3.89	3.42	2.19	.13
Third section	4.35	3.86	2.22	.14
Fourth section	5.04	4.33	2.30	.19
Total	16.98	15.30	7.26	.14

Evaluation.—This test is not markedly different from others of the paired-associates type, such as Plane Name Memory and Memory for Ships, which have correlation coefficients with Memory for Landmarks of 0.69 and 0.51 respectively. Both of these other tests yield slightly higher validity coefficients with pilot training.

Factor analysis of this test shows that 68 percent of the total variance is accounted for by common factors. Of this, 4 percent is attributed to the visual memory factor, 37 percent to the paired-associates-memory factor, and 19 percent to a third memory factor that seems to be confined to this test and Plane Name Memory. The remaining 8 percent of the total variance is accounted for by factors on which the loadings are quite low.

Since an estimate of the pilot validity, made from factorial equations (see table 28.18), is similar to that found empirically, the inference is that all factors valid for pilot selection have been accounted for in the analysis. Considerable nonerror variance, however, is still to be defined in this test.

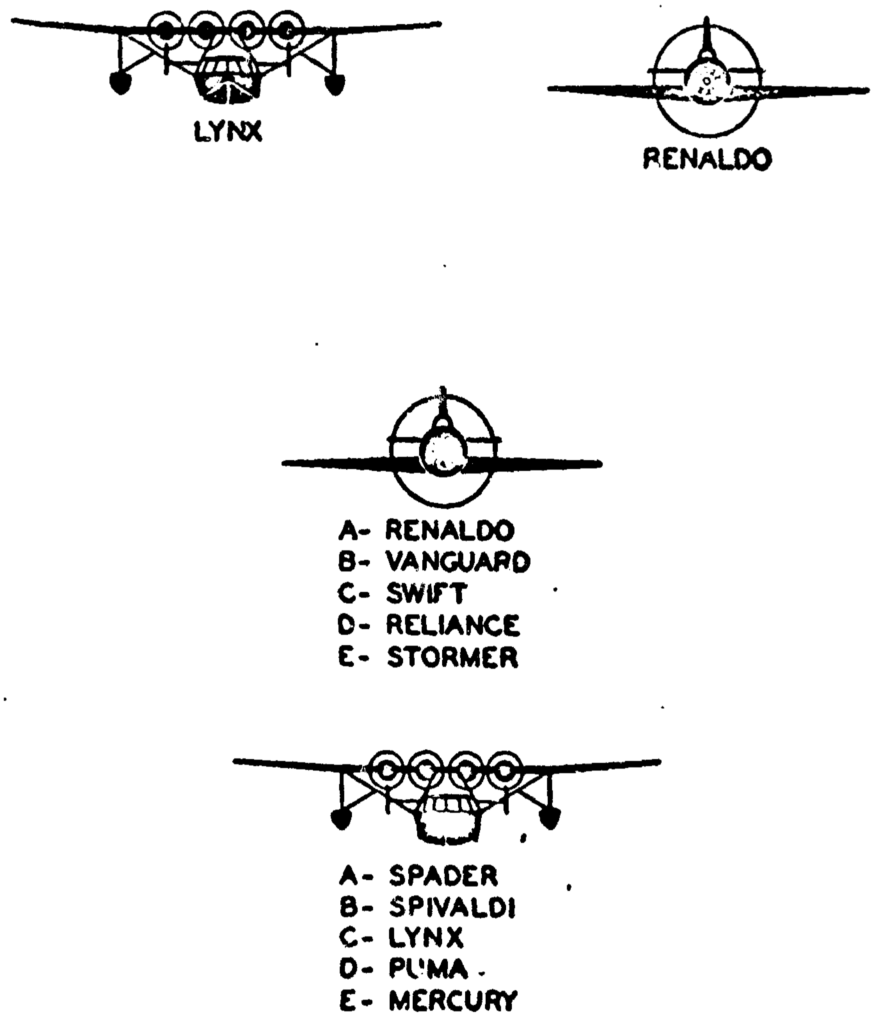


FIGURE 11.8
SAMPLE STUDY PAGE & RESPONSE ITEMS OF PLANE NAME MEMORY, CI 506AX2

This test is the purest measure developed of the paired-associates-memory factor, which accounts for 37 percent of its common variance. Both this factor and memory III have some pilot validity. Though it does not equal some of the other paired-associates memory tests for predicting pilot success, it does have something unique to offer and does have value in factor-analysis research.

Plane Name Memory, CI506AX2 [13]

This is a visual-memory, paired-associates test in which plane silhouettes are paired with their names so that later presentation of the pictorial stimulus is to call forth the verbal associate.

Description. (1) *Internal characteristics.*—The test includes 40 items which form two parts. In part I, 20 planes are shown in front-view silhouettes; in part II, 20 planes are shown in side-view silhouettes. The name of each plane appears below the silhouette. After a study period, the examinees turn to a page on which the same planes are arranged in different order with five names under each plane. The examinees are told to select the correct name of each plane. Samples of the stimulus and response items are given in the upper and lower portions respectively of figure 11.8.

(2) *Administration.*—The examinees are informed that the test is a measure of their ability to learn the names of planes. Four minutes are allowed for each study period. Six minutes are given in each part for the selection and marking of answers. The approximate total time of testing is 25 minutes.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 11.10. The distribution curves are approximately normal.

TABLE 11.10.—*Distribution constants for Plane Name Memory, CI506AX*

Group	N	M	SD
Unclassified aviation students[1]	238	20.8	8.1
Classified pilots[2]	505	23.1	8.3
Classified pilots[3]	743	29.8	6.7

[1] Tested at Psychological Research Unit No. 3, in October 1942.
[2] In class 43K. Tested at Psychological Research Unit No. 3 in January and February 1943.
[3] In class 44D. Tested at Psychological Research Unit No. 3 in September 1943.

(2) *Internal consistency.*—The degree of homogeneity of the test is indicated by a mean internal-consistency phi of 0.41, a standard deviation of the phi distribution of 0.10, and a range of values from 0.20 to 0.65. These statistics are based upon the highest 27 percent and the lowest 27 percent in total score of a group of 750 unclassified aviation students tested at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—By the alternate-forms method (part I-

[13] Developed at Psychological Research Unit No. 3. Chief contributor: Lt. Mahlon B. Smith.

part II), an estimated reliability coefficient of 0.82, corrected for length, was obtained. This figure is based on a sample of 238 unclassified aviation students tested at Psychological Research Unit No. 3 in October 1942.

(4) *Difficulty.*—Based upon item analysis of the responses of 750 unclassified aviation students tested at Psychological Research Unit No. 3, the test yielded a mean proportion of correct responses of 0.57, corrected for chance, with a range from 0.30 to 0.86, and a standard deviation of 0.16.

(5) *Factorial composition.*—The most significant loadings are in the perceptual-speed (0.29), the paired-associates-memory (0.58), and the third memory (0.51) factor. The communality is 0.71. For a fuller picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Validation results based on several samples are given in table 11.11.

TABLE 11.11.—*Validity data for Plane Name Memory, CI506AX1, using the graduation-elimination criterion*

Group	Class	Score	N_t	p_s	M_s	M_e	SD_t	r_{bis}	ρ_{bis}[1]
Pilots in primary training[2]	43G	R−W/4	222	0.74	22.12	18.79	8.16	0.24	....
Pilots in primary training[3]	43I	R−W/4	170	.86	21.06	19.38	8.19	.11	....
Pilots in primary training[4]	43J	R−W/4	131	.74	24.62	17.78	8.60	.47	....
Pilots in primary training[5]	43K	R−W/4	505	.78	23.82	20.31	8.28	.24	....
Pilots in primary training[6]	44D	R−W/4	743	.97	29.86	27.58	6.74	.15	0.18
Navigation students[7]	....	Rights[8]	1,652	.92	31.25	29.97	6.24	.10	.20
" "	....	Wrongs[8]	1,652	.92	7.35	8.61	5.50	−.11	−.20
" "	....	R−W/4[8]	1,652	.92	29.41	27.82	7.42	.11	.21

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Tested October 2 and 3, 1942, at Psychological Research Unit No. 3.
[3] Tested November 14 and 17, 1942, at Psychological Research Unit No. 3.
[4] Testing dates and locale not available.
[5] Tested January 20, February 8 and 9, 1943, at Psychological Research Unit No. 3.
[6] Tested September 2 and 3, 1943, at Psychological Research Unit No. 3.
[7] Tested May 31 and June 1, 1944, at Psychological Research Unit No. 1; April 17 through 21, 1944, at Psychological Research Unit No. 2; and April 10, 11, and 12, 1944, at Psychological Research Unit No. 3.
[8] For this sample, the correlation between rights and wrongs is −.83.

Evaluation.—Plane Name Memory, another of the paired-associates type of test, shows relatively moderate pilot and navigator validity (approximately 0.22 and 0.21, respectively). Factor analysis of this test shows that 71 percent of the total variance is accounted for by common factors, compared with a reliability of 0.82. Of this, the perceptual-speed factor accounts for 8 percent, the paired-associates-memory factor for 34 percent, and a third memory factor for 26 percent. This latter factor seems to be restricted to this test and Memory for Landmarks and will be discussed later in the chapter. The remaining 3 percent of the known variance is accounted for by factors on which the loadings are quite low.

An estimation of the pilot validity coefficient of this test, made from factors (see table 28.17) for which the pilot validity is known, accounts for approximately 0.16 of the validity coefficient of 0.22. The difference may be due to variance in visualization, which was lacking in the particular analysis in which this test appeared.

Memory for Ships, CI504AX [18]

This is another visual-memory test in which a pictorial symbol and a verbal symbol are paired so that the pictorial stimulus is to stimulate recall of the verbal associate.

Description (1) *Internal characteristics.*—The test consists of three study pages and three response pages. On each study page are 10 ships paired with their respective nationalities. Succeeding the study page is a response page on which 12 ships are presented, without their nationalities indicated, 10 of which are the same as on the previous page, and 2 that are not. The ships are shown from an oblique aerial view and all are headed in the same direction..

The task of the examinee is to determine the nationality of each ship as shown on the study page, or, if it did not appear on the study page to indicate that it was not shown. At the top of each response page is placed a letter symbol for each nationality which is used in marking the answers. Figure 11.9 shows a portion of both the stimulus and the response pages.

(2) *Administration.*—The examinees are informed that the test is to see how well they can remember ships and their nationalities. Two minutes are allowed to study the ships; then, at a signal, the page is turned and 3 minutes are allowed for answering the problems. The total testing time is approximately 30 minutes.

(3) *Scoring.*—The scoring formula is R—W/5.

Statistical results.—All the data following are for examinees at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—A sample of 238 unclassified aviation students tested in November 1942 yielded a mean score of 15.9 and a standard deviation of 6.4. The distribution curve is approximately symmetrical and flatter than normal.

(2) *Reliability coefficient.*—Correlations among the three parts of this test yielded the estimates of reliability given in table 11.12.

TABLE 11.12.—*Estimated alternate-forms reliability coefficients for Memory for Ships, CI504AX based upon a sample of 238 unclassified aviation students*

Parts	r_{tt} [1]	r_{tt} [2]
Part I with Part II	0.34	0.61
Part I with Part III	.38	.65
Part II with Part III	.48	.74

[1] Tested in November 1942.
[2] Corrected for triple length.

(3) *Factorial composition.*—The most significant loadings are in the paired-associates-memory (0.50), spatial-relations (0.31), and the per-

[18] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lian Hutchinson, Lt. David H. Jenkins.

ceptual-speed (0.29) factors. The communality is 0.51. For a fuller picture of the factorial composition of this test, see appendix B.

(4) *Test validity.*—Validation results based on two samples are given in table 11.13.

TABLE 11.13.—*Validity data for Memory for Ships, CI504AX, based upon the graduation-elimination criterion in primary pilot training*

N_t	p_g	M_g	M_e	SD_t	r_{bis}
[1]658	0.73	16.14	14.08	6.30	0.20
[2]586	.79	17.71	15.88	6.27	.17

[1] In class 43I. Tested November 9, 10, 11 and 17, 1942.
[2] Eighty-nine cases in class 43I, 154 cases in class 43J, and 343 cases in class 43K. Tested November 14, 1942, January 20, 1943, and February 8, 9, and 10, 1943.

Evaluation.—Memory for Ships has a moderately low validity for the prediction of pilot success (average validity coefficient of 0.17), but a sufficiently low correlation with the pilot stanine so that it would add, though by an amount rarely worth considering, to the validity coefficient if used in conjunction with the classification battery.

Factor analysis of this test shows that 51 percent of the total variance is accounted for by common factors, leaving a fair amount of undefined nonerror variance. Of the known variance, the perceptual-speed factor accounts for 8 percent, the spatial-relations factor for 10 percent, and the paired-associates-memory factor for 25 percent. The remaining 8 percent of the total variance is accounted for by factors on which the loadings are quite low. That the spatial-relations factor accounts for 10 percent of the common variance of this test seems unusual, but the fact that, in memory of naval ships, spatially complicated structures are involved seems reasonable rationalization for it.

An estimate of the pilot validity for this test (see table 28.18) is 0.20, which is nearly the same as the empirical value (0.17). This indicates that all factors valid for the pilot have been accounted for. This test, like Memory for Plane Silhouettes, has relatively high loadings on the perceptual and spatial-relations factors and thus is not as pure a measure of the paired-associates factor as is Memory for Landmarks.

SYMBOLIC MEMORY TESTS

The second main category under which memory tests are grouped is the memory for verbal, or more accurately, symbolic material. Here also the tests falling in this classification may be divided into those that involve the ability to remember and to recognize complex wholes and relations of parts, and those that require the remembering of simple wholes (paired associates). The former includes tests in which the stimulus is auditory and those in which the stimulus is printed. The latter group as represented here consists only of tests with a verbal (printed) stimulus and response.

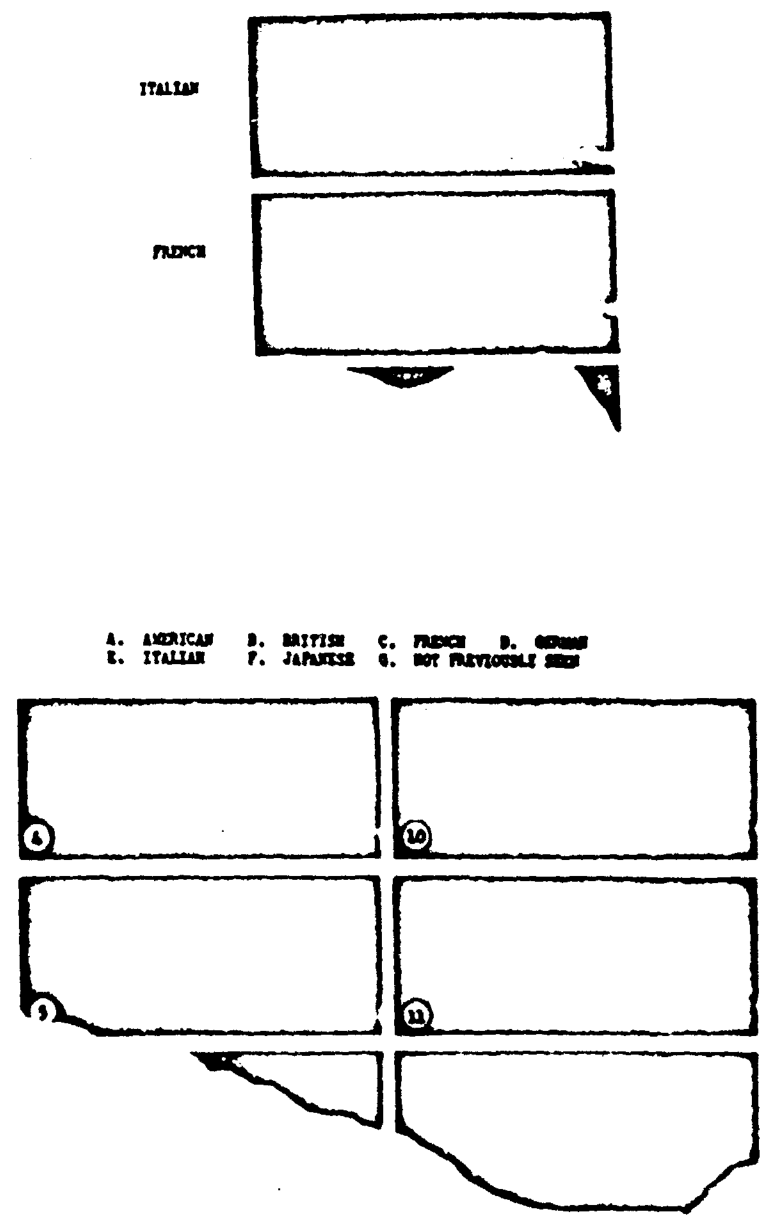


FIGURE 11.9
PORTIONS OF A STUDY PAGE AND TEST ITEMS OF MEMORY FOR SHIPS, CI504AXI

Rationale

Memory for symbolic material is more generally involved in learning and in training, and is thus perhaps more important than memory for pictorial material. The general rationale behind this area hardly needs to be elaborated and is covered in the earlier part of this chapter. The necessity for this type of memory does not cease upon completion of training, but continues throughout combat. Briefing for a combat mission involves not only pictorial material, but also verbal material, which supplements the pictorial material. Much of this material is presented orally as well as in printed form. The auditory presentation of complex material is represented by the Memory for Tactical Plans test, while the written presentation of complex material is represented by the Geographical Memory Test. Memory for simple symbols (paired associates) is represented by the Memory for Plane Designation test, which measures a basic type of learning and memory.

Memory for Tactical Plans, CI509BX [14]

This is a verbal, auditory-memory test designed to measure ability to remember meaningful material (instructions) over a longer term than used in immediate-memory tests. The stimulus is presented auditorily, and the response is made to printed questions. There are three closely similar forms of this test. The original form, CI509AX, was subjected to item analysis and revised to provide form CI509BX. With but extremely slight changes, this form was phonographically recorded and called form CI509C.

Description. (1) *Internal characteristics.*—The examinees are read a summary of briefing data for a mock bombing mission. About 2 hours later, after other tests have been interposed, 40 simple memory questions, divided into two parts of 20 items each, concerning the briefing data are asked. A sample paragraph of the briefing data and corresponding items are as follows:

Major Carpenter's flight will follow 4 minutes behind Major Wilson's flight at an altitude of 21,000 feet. They will carry 500-pound bombs and incendiaries. Major Carpenter's flight will have the additional assignment of photographing the bombed area.

Sample items are as follows:

Major Carpenter's flight will carry:

A. 100-pound bombs and incendiary bombs.
B. 500-pound bombs and incendiary bombs.
C. 1000-pound bombs and incendiary bombs.
D. 2000-pound bombs and incendiary bombs.
E. Block busters.

Major Carpenter's flight has the assignment of:

A. Attacking the troop loading zone.
B. Attacking the roundhouse.
C. Attacking the warehouse.

[14] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Milton Burdman, Capt. Harry Rosenberg.

D. Photographing the bombed area.
E. Bombing the alternative objective.

(2) *Administration.*—The pertinent parts of the directions preceding the briefing are as follows:

. . . . later in the day you will be asked to answer questions based upon what you hear now.

Assume that you are a member of a flight which is to take part in a bombing raid. You are listening to the instructions of your flight commander.

The interim between the briefing and the written questions varied from 2 to 3 hours among the different forms. The directions and briefing for CI509C were phonographically recorded in an effort to achieve greater standardization. Total testing time is approximately 25 minutes, with 10 minutes allowed for the directions and briefing.

(3) *Scoring.*—The scoring formula is R—W/4 in all forms.

Statistical results.—All the data given below are for examinees at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 11.14. The distribution curves are approximately normal.

TABLE 11.14.—*Distribution constants for Memory for Tactical Plans, CI509*

Group	Form	N	M	SD
Unclassified aviation students[1]	CI509AX	365	16.9	6.9
Classified pilots[2]	CI509BX	790	25.7	6.4
Classified pilots[3]	CI509BX	570	21.8	7.0

[1] Tested October 23 and 24, 1942.
[2] In class 44D. Testing dates not reported.
[3] In class 43J. Tested January 9, 11, 13, 15, and 16, 1943.

(2) *Internal consistency.*—The degree of homogeneity of the items (in form CI509BX) is indicated by a mean internal-consistency phi of 0.36, a standard deviation of the phi distribution of 0.12, and a range of values from 0.02 to 0.59. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 700 unclassified aviation students.

(3) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.68, corrected for length, was obtained for the AX form. This figure is based on a sample of 500 unclassified aviation students tested in October 1942.

(4) *Difficulty.*—Based upon item analysis of the responses of 700 unclassified aviation students, the test (form CI509BX) yielded a mean proportion of correct responses of 0.54, corrected for chance, with a range from 0.08 to 0.90 and a standard deviation of 0.19.

(5) *Factorial composition.*—The most significant loadings of the AX form are in the verbal (0.57) and visualization (0.32) factors. The

communality is only 0.47. For a fuller picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Validation results based on several samples are given in table 11.15.

TABLE 11.15.—*Validity data for Memory for Tactical Plans, CI509BX (graduation-elimination used as criterion in all samples)*

Group	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training[2]	570	0.84	22.15	19.96	7.05	0.17	0.20
Pilots in basic training[2]	477	.91	22.18	21.67	6.81	.04	.16
Pilots through basic training[2]	570	.76	22.18	20.53	7.05	.14	.17
Pilots in primary training[3]	790	.93	25.93	23.29	6.45	.20	.19

[1] Assuming an unrestricted stanine standard deviation of 1.09.
[2] Same sample followed through primary and basic training in class 43J. Tested January 1943.
[3] In class 44D. Testing dates not reported.

Evaluation.—Since the correlation of this test with the pilot stanine is low, in spite of rather low validity, it would be of some value to the classification battery. Other features, however, such as the difficulty of administration, made impractical its inclusion in the battery.

Since this test was designed to measure a relatively long-term type of memory, it is unique among the memory tests. It does not appear to be loaded with any of the memory factors in common with short-term memory tests. Factor analysis shows that only 47 percent of the total variance of the test is accounted for by common factors, compared with a reliability of 0.68. Of this, the verbal factor accounts for 32 percent and the visualization factor for 10 percent. This is to be expected, since the material is presented orally and tested with questions involving the relative positions of three flights attacking a target. The remaining 5 percent of the total variance accounted for by the analysis is on factors on which the loadings are quite low.

Examination of the duties of air crew indicates that this type of memory should be important. Since this is the only test measuring more than immediate memory, and since it has validity, further development would be worth while. A study of its unknown valid factor, or factors, would be profitable, for, when this variance is properly identified, a more unique test without verbal variance might be constructed. Its validity for pilot selection is very largely unaccounted for by known factors. Its average obtained validity is 0.19, whereas that expected from known factors is only 0.06. Between these two values lies rich unexplored territory.

Geographical Memory, CI508AX[18]

This is a symbolic memory test involving the relationships of parts to

[18] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, Lt. Lina Hutchinson.

a complex whole. Both the stimulus and the response part of the problems are presented in printed form.

Description. (1) *Internal characteristics.*—In this test, written descriptions of geographical areas, each approximately a typewritten page long, are presented for the examinees to study. At the end of each study period, questions are asked about the location of important points, their direction and distance from each other, and details about transportation routes. Answers to some of the questions are not specifically stated in the description, but can be determined from the information given. The test consists of descriptions of two geographical areas with a cluster of 20 items concerning the first, and 25 items concerning the second description. A paragraph from the geographical description and corresponding item are presented below.

The Olson-Van Ruyan Marine Engine Corporation factory is near the northwest corner of the bay. It is served by a single-track railroad, coming from the north. A two-lane highway runs along the north shore of the bay from the Olson-Van Ruyan factory to Warrenton at the northeast corner of the bay. Warrenton also extends along the east bay shore for about 4 miles. Commercial docks extend for about 3 miles midway on the east bay shore. An oil pipe line, bringing oil from wells farther east, terminates at the docks. Near the east end of the south shore of the bay is the Great Western Shipyard, and near the western end is the Admiralty Seaplane Base.

The distance from the Olson-Van Ruyan factory to the seaplane base is about:

A. 2 miles.
B. 5 miles.
C. 8 miles.
D. 11 miles.
E. 15 miles.

Goods arriving at the Olson-Van Ruyan Marine Engine factory from the north would likely come by:

A. Truck over the 4-lane highway.
B. Single track railroad.
C. Double track railroad.
D. Truck over the 2-lane highway.
E. Truck over the 3-lane highway.

(2) *Administration.*—The directions for the test are as follows:

This is a test of your ability to remember details of a geographical description. You will have 7 minutes to study a written description of a geographic area. At the end of that time you will be asked questions based upon the description. You may find it helpful to imagine a map of the area described.

You should note especially the following characteristics of the area.

1. Location of important details.
2. Direction of important points from each other.
3. Distance of important points from each other.
4. Location and details of transportation routes . . .

As indicated above, 7 minutes are allowed for study of the description on each part. Ten minutes are given for selection and marking of answers on the first part and 12 minutes for the second part. The total testing time is 40 minutes.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data are for examinees tested at Psychological Research Unit No. 3 on October 19 and 20, 1942.

(1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 11.16. The distribution curves are approximately symmetrical and somewhat flatter than normal.

TABLE 11.16.—*Distribution constants for Geographical Memory, CI508AX*

Group	Part	N	M	SD
Unclassified aviation students	Total	185	12.9	9.2
Classified pilots	I	223	5.7	4.5
Classified pilots	II	223	8.4	6.1

(2) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.74, corrected for length, was obtained. This figure is based on a sample of 250 unclassified aviation students.

(3) *Test validity.*—Validation results based on several samples are given in table 11.17.

TABLE 11.17.—*Validity data for Geographical Memory, CI508AX, based upon a sample of 223 pilots (graduation-elimination in primary training used as criterion; p_s=.75)*

Part	M_p	M_e	SD_t	r_{bis}
I	5.85	5.37	4.54	0.06
II	8.67	7.37	6.12	.33

Evaluation.—This test had been intended to resemble Map Memory, CI505AX1, in all respects except that the geographical material was presented in verbal rather than visual form. Presenting the material verbally caused a high correlation with the verbal test, Reading Comprehension (0.48), which is in the classification battery. This would mean a very high loading of the verbal factor in this test, possibly as high as 0.70. This fact probably indicates a lack of factorial resemblance to Map Memory, and so it was felt further development of this test is not worth while.

Memory for Plane Designations, CI507AX [16]

This is a paired-associates test utilizing symbolic material in literal and verbal form for the stimulus and the response respectively.

Description. (1) *Internal characteristics.*—The test consists of 2 parts of 20 items each. In each part there is a study page on which the names of 20 hypothetical airplane manufacturers are given, each paired with a three-letter symbol, somewhat like those given by the Navy for plane designation. For example:

P-YD .... O'Rourke.
P-LC .... Inman.
P-ZI .... Brennerman.

The first letter is the same throughout the part, but no combinations of the last two letters are similar for different associates. After a study period, the examinees turn to the response-page on which the 20 symbols are given with five names listed below each one. A sample item follows:

P-LC
- *a.* Dalton
- *b.* O'Rourke
- *c.* Brennerman
- *d.* Inman
- *e.* Powers

The task of the examinees is to select from the five choices that name that has been paired with the symbol on the study page. A later form, CI507BX, is similar, except the symbols are paired with plane names instead of the names of manufacturers.

(2) *Administration.*—Six minutes are allowed for the study period and the same for selecting and marking the answers. The testing time is approximately 25 minutes.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results—The data given below are for examinees tested at Psychological Research Unit No. 3 on October 13, 15, and 17, 1942.

(1) *Distribution statistics.*—Typical examples of distribution statistics obtained on CI507BX, the later form of this test mentioned above, are given in table 11.18. The distribution curves are approximately symmetrical and somewhat flatter than normal.

TABLE 11.18.—*Distribution constants for Memory for Plane Designations, CI507BX based upon a sample of 185 unclassified aviation students*

Part	M	SD
I	6.8	4.5
II	7.1	7.8
Total	13.9	4.8

[16] Developed at Psychological Research Unit No. 3. Chief contributor: Capt. Harry Rosenberg.

(2) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.82, corrected for length, was obtained. This figure is based on a sample of 367 unclassified aviation students.

(3) *Test validity.*—A sample of 348 pilots yielded a biserial correlation of −0.03 between performance in this test and the graduation-elimination criterion in primary pilot training. The mean score for graduates was 8.33, for eliminees 8.56, and the standard deviation for both combined was 5.18. Of this sample, 67 percent were graduates.

Evaluation.—Memory for Plane Designation does not correlate to any degree with any of the classification tests, the highest correlation being with Reading Comprehension (0.24) and with Numerical Operations (0.24). This test would be expected to have some degree of relationship with the former test because of the emphasis on verbal symbolism in both of them. The low pilot validity coefficient of this test is consistent with its verbal and numerical content. Although it has not been factor analyzed, its zero validity leads us to expect no significant variance in any valid factor. Because of these considerations, no further development of the test was undertaken.

FACTOR ANALYSIS OF MEMORY TESTS [*]

In order to obtain a clearer picture of the memory area and of the memory tests developed in the aviation-psychology program, two factor analyses were made of all the memory tests that were ready in the summer of 1942.

The Data

Two batteries were administered in order to obtain the intercorrelations for use in the analysis. Twelve tests appear in battery I, five of which were memory tests and the remaining were tests selected from the classification battery. The second battery included 13 tests, 6 of which were memory tests and the remaining were the same classification tests used in battery I. Thurstone's centroid method with rotation of axes was employed. The sample for battery I was composed of 179 classified bombardiers, and for battery II of 238 unclassified aviation students. In both cases the range of talent was practically unrestricted by selection other than that produced by the AAF Qualifying Examination. In both correlation matrices, the intercorrelations of classification tests were based on 527 unclassified students, assuming that they were comparable to the two special samples and with the belief that much more stable data were thus obtained. The intercorrelations for the two batteries are presented in tables 11.19 and 11.20, the centroid loadings in tables 11.21 and 11.22, and the rotated factor loadings in tables 11.23 and 11.24.

[*] Executed by S/Sgt. J. Gordon Ellis, S/Sgt. Benjamin Fruchter, Capt. Lloyd G. Humphreys, and Sgt. Harold H. Singer.

703130—47—18

TABLE 11.19.—*Correlation matrix for Memory Battery I*[a]

Test	1	2	3	4	5	6	7	8	9	10	11	12
1. Speed of Identification[b]	....	47	40	22	20	11	−01	26	22	40	12	06
2. Spatial Orientation I	[a]47	...	48	23	21	17	14	39	29	31	14	03
3. Spatial Orientation II	40	48	..	23	30	21	13	55	34	41	16	23
4. SAM Complex Coordination	22	23	23	..	29	21	17	30	27	31	10	17
5. Mechanical Comprehension, AC10D, Pt. VI	20	21	30	29	..	33	27	29	32	25	04	27
6. Reading Comprehension, AC10D, Pt. II	11	17	21	21	33	..	47	57	37	15	35	44
7. Arithmetic Reasoning	−01	14	13	17	27	47	....	32	28	12	21	28
8. Map Memory, CI505AX1	26	39	55	30	29	57	32	..	61	25	32	24
11. Map Memory, CI505AX3	22	29	34	27	32	37	28	61	..	26	29	29
12. Memory for Plane Silhouettes	40	31	41	31	25	15	12	25	26	..	41	11
13. Memory for Landmarks	12	14	16	10	04	35	21	32	29	41	..	12
14. Memory for Tactical Plans	06	03	23	17	27	44	28	24	29	11	12	..

[1] Decimal points have been omitted.
[2] For code numbers see table 11.21.
[a] It should be noted that the correlations among variables 1–7 are identical in tables 11.19 and 11.20. These were computed for a sample of 527 unclassified aviation students which does not include those taking the memory tests. The latter group was a sample of 179 classified bombardiers.

TABLE 11.20.—*Correlation matrix for Memory Battery II*[a]

Test	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Speed of Identification[b]	....	47	40	22	20	11	−01	24	29	41	15	15	27
2. Spatial Orientation I	[a]47	...	48	23	21	17	14	28	37	23	21	22	25
3. Spatial Orientation II	40	48	..	23	30	21	13	25	41	32	15	32	27
4. SAM Complex Coordination	22	23	23	..	29	21	17	13	22	21	12	15	28
5. Mechanical Comprehension, AC10D, Pt. VI	20	21	30	29	..	33	27	17	35	26	13	25	15
6. Reading Comprehension, AC10D, Pt. II	11	17	21	21	33	..	47	31	33	35	28	28	21
7. Arithmetic Reasoning	−01	14	13	17	27	47	....	29	24	−02	26	15	10
9. Map Memory, CI505BX1	24	28	25	13	17	31	29	..	55	35	44	35	28
10. Map Memory, CI505AX2	29	37	41	22	35	33	24	55	..	44	37	28	30
12. Memory for Plane Silhouettes	41	23	32	21	26	35	−02	35	44	....	34	44	46
13. Memory for Landmarks	15	21	15	12	13	28	26	44	37	34	..	69	51
15. Plane Name Memory	15	22	32	15	25	28	15	35	28	44	69	..	47
16. Memory for Ships	27	25	27	28	15	21	10	28	30	46	51	47	..

[1] Decimal points have been omitted.
[2] For code numbers see table 11.22.
[a] See footnote to table 11.19. The sample for the memory tests was composed of 238 unclassified aviation students.

Table 11.21.—*Centroid factor loadings and communalities for Memory Battery I*[a]

Test	I	II	III	IV	V	VI	VIII	h^2
1. Speed of Identification, CP610A	46	−47	−10	−07	−09	−13	−09	48
2. Spatial Orientation I, CP501B	52	−36	−15	−11	−22	09	09	50
3. Spatial Orientation II, CP503B	53	−28	−14	−20	08	−07	18	58
4. Complex Coordination, CM701A	44	−15	18	10	10	14	−15	31
5. Mechanical Comprehension, AC10D	49	−05	32	−07	15	07	−09	38
6. Reading Comprehension, AC10D	62	45	14	−04	−26	−13	−19	73
7. Arithmetic Reasoning, CI206B	45	35	25	12	−16	12	11	44
8. Map Memory, CI505AX1	74	14	−19	−24	−06	12	−11	69
11. Map Memory, CI505AX3	65	19	−17	−13	18	25	−13	62
12. Memory for Plane Silhouettes, CI503AX1	53	−29	−17	41	16	−19	−11	64
13. Memory for Landmarks, CI510AX1	42	20	−30	34	−08	−19	−14	48
14. Memory for Tactical Plans, CI509AX1	42	−26	30	−12	[illegible]	−29	09	47

[a] Decimal points have been omitted.

Table 11.22.—*Centroid factor loadings and communalities for Memory Battery II*[a]

Test	I	II	III	IV	V	VI	VII	VIII	h^2
1. Speed of Identification, CP610A	48	43	−14	−11	04	−16	21	−12	53
2. Spatial Orientation I, CP501B	53	31	05	−25	−13	−16	19	08	53
3. Spatial Orientation II, CP503B	56	34	06	−13	−15	14	12	10	52
4. Complex Coordination, CM701A	41	18	−06	33	−13	−14	−13	07	37
5. Mechanical Comprehension, AC10D	46	15	15	23	−09	19	−08	−13	38
6. Reading Comprehension, AC10D	53	−16	26	33	07	08	19	−14	55
7. Arithmetic Reasoning, CI206B	38	−31	48	22	−18	−19	07	−12	60
9. Map Memory I, CI505BX1	60	−15	16	−23	29	−10	−12	−09	58
10. Map Memory II, CI505AXII	67	09	23	−13	20	04	−20	−05	61
12. Memory for Plane Silhouettes, CI503AX1	62	08	−35	10	37	19	08	−09	70
13. Memory for Landmarks, CI510AX1	62	−52	−25	−23	−06	−06	−12	−19	83
15. Plane Name Memory, CI506AX1	63	−34	−29	−12	−18	26	−03	−05	71
16. Memory for Ships, CI504AX1	58	−10	−39	03	−03	−11	−03	05	52

[a] Decimal points have been omitted.

TABLE 11.23.—*Rotated factor loadings for Memory Battery I*[1]

Test	I	II	III	IV	V	VI	VII	h^2
1. Speed of Identification[2]	64	14	02	21	−03	−01	−01	47
2. Spatial Orientation I	61	04	−04	28	−07	13	18	51
3. Spatial Orientation II	61	16	07	09	−01	12	38	57
4. SAM Complex Coordination	22	13	06	46	16	06	11	32
5. Mechanical Comprehension	14	27	−05	36	27	−02	29	38
6. Reading Comprehension	10	67	21	12	39	24	02	73
7. Arithmetic Reasoning	−05	35	20	04	47	22	07	44
8. Map Memory, CI505AX1	35	42	09	16	06	54	26	70
11. Map Memory, CI505AX3	18	31	15	21	14	55	31	61
12. Memory for Plane Silhouettes	43	07	50	43	−01	−07	02	63
13. Memory for Landmarks	14	25	56	12	−04	21	−13	47
14. Memory for Tactical Plans	−02	57	10	08	09	−12	32	47

[1] Decimal points have been omitted.
[2] For code numbers see table 11.21.

TABLE 11.24.—*Rotated factor loadings for Memory Battery II*[1]

Test	I	II	III	IV	V	VI	VII	VIII	h^2
1. Speed of Identification[2]	66	−04	08	17	−04	22	09	−09	539
2. Spatial Orientation I	62	−04	24	07	18	20	00	11	533
3. Spatial Orientation II	63	25	11	09	02	11	08	06	502
4. SAM Complex Coordination	22	06	03	52	14	08	07	12	369
5. Mechanical Comprehension	23	36	07	33	[illegible]	10	15	−14	361
6. Reading Comprehension	−01	52	18	26	40	12	−03	00	546
7. Arithmetic Reasoning	−05	27	−02	15	68	09	14	04	590
9. Map Memory, CI505BX1	22	05	41	08	23	52	05	07	556
10. Map Memory, CI505AX2	35	23	14	16	17	58	07	−02	591
12. Memory for Plane Silhouettes	25	26	62	32	−11	20	−09	−04	679
13. Memory for Landmarks	17	−11	66	02	33	19	44	−09	824
15. Plane Name Memory	29	13	58	06	10	02	51	−04	713
16. Memory for Ships	29	−08	50	31	10	06	20	13	507

[1] Decimal points have been removed.
[2] For code numbers see table 11.22.

Results

The results of the analysis of the two batteries will be summarized together, because many of the tests overlap and identical factors were extracted from the two sets of intercorrelations. A test is listed under a factor if a loading of 0.30 or higher is attained in either analysis.

Rotated factor I is the well-verified perceptual-speed factor in that the Speed of Identification and the Spatial Orientation I and II tests appear with by far the greatest loadings. Test loadings on this factor are as follows:

Test number	Test name	Loadings	
		I	II
1	Speed of Identification, CP610A	0.64	0.66
2	Spatial Orientation I, CP501B	.61	.62
3	Spatial Orientation II, CP503B	.61	.63
12	Memory for Plane Silhouettes, CI503AX1	.43	.25
8	Map Memory, CI505AX1	.35	(1)
10	Map Memory, CI505AX2	[1]....	.35
15	Plane Name Memory, CI505AX1	....	.29
16	Memory for Ships, CI504AX1	....	.29

[1] For this and the following factors, blanks indicate the absence of a test in a particular analysis.

Rotated factor II has significant loadings in the following tests:

Test number	Test name	Loadings	
		I	II
6	Reading Comprehension, AC10D	0.67	0.53
14	Memory for Tactical Plans, CI509AX1	.57	....
8	Map Memory, CI509AX1	.42	....
7	Arithmetic Reasoning, CI206B	.35	.27
11	Map Memory, CI505AX3	.31	....
5	Mechanical Comprehension, AC10D	.27	.36

This is obviously the verbal factor found in most analyses. While there are certain variations in the loadings of memory tests on this factor, it seems clear that of these tests Memory for Tactical Plans is the most verbal. The map-memory tests are next most heavily loaded with verbal comprehension. The average verbal loadings of the other memory tests are not appreciable.

The following data define rotated factor III:

Test number	Test name	Loading	
		I	II
12	Memory for Plane Silhouettes, CI503AX1	0.50	0.62
13	Memory for Landmarks, CI510AX1	.56	.66
15	Plane Name Memory, CI506AX1	....	.58
16	Memory for Ships, CI505AX1	....	.50
9	Map Memory, CI505BX1	....	.41

The tests high on this factor are, with one exception, fundamentally of the paired-associates form. Whether this factor should be so defined, that is, in terms of the form of memorizing, is uncertain. It is perhaps broader than this and could, on the basis of present evidence, be called a "rote-memory" or an "associative-memory" factor. All but one of the map-memory tests have near zero loadings here. This one, CI505BX1, is heavily weighted with items based on a schematic map. The detail on this map consists of names of cities, connecting roads, and mileages. Such material could be expected to introduce a high "rote-memory" loading, hence the assumption of a general rote-memory factor receives some support.

The hypothesis of an associative-memory factor, however, better accounts for the clean-cut distinction between the two groups of memory tests—the recall by association on the one hand and the recall by reproduction on the other. This hypothesis needs the support of other evidence—finding the same factor in tests of serial learning, for example. For lack of more crucial evidence, it seems best to call this reference variable the paired-associates memory factor, staying close to the more apparent characteristics of the tests defining it.

Rotated factor IV is defined by the following data:

Test number	Test name	Loadings I	Loadings II
4	Complex Coordination, CM701A	*0.46	0.52
12	Memory for Plane Silhouettes, CI503AX2	.43	.32
5	Mechanical Comprehension, AC10C	.36	.33
16	Memory for Ships, CI504AX1	....	.31

This is undoubtedly the spatial-relations factor, which in this and other analyses has been defined by the Complex Coordination test. Memory for Plane Silhouettes and Memory for Ships have moderate loadings with this factor. It is evident that these tests are so constructed that persons high on the spatial-relations factor are aided in memorization of this material.

The following tests have significant loadings on rotated factor V:

Test number	Test name	Loadings I	Loadings II
7	Arithmetic Reasoning, CI206B	0.47	0.68
6	Reading Comprehension, AC10D	.39	.40
13	Memory for Landmarks, CI510AX1	−.04	.33

This is most probably the general-reasoning factor always found in Arithmetic Reasoning and Reading Comprehension. The discrepancy in loading for Memory for Landmarks is quite unusual. In view of the absence of this factor in other memory tests, one is led to suspect that the zero loading here is more nearly correct.

Rotated factor VI is defined by the following data:

Test number	Test name	Loadings I	Loadings II
8	Map Memory, CI505AX1	0.54	
11	Map Memory, CI505AX3	.55	
9	Map Memory, CI505BX1		0.52
10	Map Memory, CI505AX2		.58

This factor is restricted to map-memory tests in these analyses. It is possible, however, that the factor is more general than this. This factor has been called visual memory in view of the obvious visual content of the tests that define it. It could be hypothesized that this is a more general reproductive-memory variable, but evidence is lacking as to whether it would be held in common by auditory and other types of memory tests. Following these analyses, additional tests such as Plane Formation Memory, CI513A, and Plane Position Memory, CI512A, were constructed for the purpose of purifying tests for the supposed visual-memory factor and for further clarification of this hypothesis by analysis.

Rotated factor VII is defined by the following data from battery I only:

Test number	Test name	Loading	
		I	II
3	Spatial Orientation II, CP503B	0.38	....
14	Memory for Tactical Plans, CI509AX1	.32	....
11	Map Memory, CI505AX3	.31	....
5	Mechanical Comprehension, AC101D	.29	....

Though this factor is not well-defined, it is probably visualization. Other analyses have shown Spatial Orientation II, CP503BX, to have a moderate loading in this factor. Most mechanical-comprehension tests have moderate to high visualization loadings. Although Memory for Tactical Plans involves delayed memory for auditorially presented verbal material, it is easy to rationalize a visualization component in the test. Comprehension of the described mission and the answering of verbal questions about it could both involve some visualization to good advantage.

It is of considerable psychological importance to find that there is a clear-cut hiatus between reproductive visualization or visual memory (factor VI) and another type of visualizing which appears to require manipulation and so may be called manipulative visualization. More discussion of this point will be found in the chapter on visualization (ch. 12).

Rotated factor VIII is defined by the following data from battery II only:

Test number	Test name	Loading	
		I	II
15	Plane-Name Memory, CI506AX1	0.51	....
13	Memory for Landmarks, CI510AX1	.44	....

This factor seemingly constitutes an unimportant doublet of the two memory tests that are most similar to each other. Except for differences in subject matter, these two tests are practically identical measures of the same basic functions. The correlation between the two tests is 0.69. Both involve the pairing of relatively simple figure and verbal symbols. It is not unreasonable to speculate that in addition to a general paired-associates memory ability (factor III above) common to all tests in which items of information are memorized by pairs, there is also for each type of pair a more restricted ability to learn, retain, and recall.

Should there also be an associative-memory factor, even more general than that for memorizing in pairs (paired-associate factor), a complex hierarchy of memory abilities would exist with separable variables of different degrees of generality. The structure of memory abilities is thus seen to need a thorough investigation from the factorial approach.

Rotated factor IX in memory battery II is seemingly a true residual factor. The ninth centroid was used in the rotations of the other eight. When the rotations were completed, factor IX had a smaller share of the total variance in the battery than it had before rotations were started. Since the centroid method does not extract the maximum common variance possible with each successive factor, such an outcome is not unexpected.

Conclusions

In conclusion, it can be definitely stated that two general memory factors have been isolated. A third memory factor specific to a particular type of test has also been isolated. It may be that memory factors of limited scope can be multiplied almost indefinitely by relatively slight changes in the format or content of memory tests. Such factors might be useful in prediction within their own areas.

It would seem that at least one important memory factor remains to be described. Memory for Tactical Plans, CI509, a test of delayed memory for orally-presented material, does not appear to be weighted in either of the two general memory factors. It is quite possible that addition of similar delayed-memory tests to these batteries would result in the isolation of a delayed-memory factor of some kind, or, instead, one peculiar to auditory-verbal presentation, or both.

BIBLIOGRAPHY

(1) H. B. Carlson, Factor analysis of memory ability, *J. Exp. Psychol.*, 1937, 5, 477–492.

CHAPTER TWELVE

Visualization Tests[1]

INTRODUCTION

Historical Background

One of the earliest systematic attempts to investigate the field of visualization was a study conducted over 60 years ago by Sir Francis Galton. His questionnaire concerning the vividness with which objects from the morning's breakfast table could be imagined, or visualized, is a famous pioneer study. The field later proved interesting and challenging to many investigators. At one time it even became the battlefield for a fundamental psychological controversy as to whether there is any thought process at all that does not involve some kind of images, visual or otherwise.

Practical difficulties have retarded experimentation in this field. Visualizing is a private experience, involving little or no overt behavior that can be measured. It has been extremely difficult, therefore, to devise objective experiments even to prove unequivocally that visualization exists. Such experiences as dreams, hallucinations, vivid memories, eidetic images, and so on are common enough, however, to convince most observers that some process of visual imagination or visualization does exist, even though it defies most attempts at objective measurement.

The older attempts at quantifying and measuring visualization were aimed chiefly at discovering the degree of clarity or sharpness, the persistence, and the frequency with which visualization took place. More or less implicitly, these studies assumed that if visualization occurs at all, the process is always the same; that is, the ability to visualize was assumed to be a single unitary ability rather than a family of related abilities. Though this assumption is natural and understandable, it is certainly not the only one that can be made. It may be that there are several abilities to visualize and that different tasks require different visualizing abilities. More recently, some of the psychological problems involved have been enumerated as follows:

> Is visualizing flat forms the same as visualizing solid forms as they would appear from different sides? Is visualizing solid objects the same as visualizing movement of parts in a diagram of a machine? Considering only flat forms, is the same ability required in visualizing several shapes singly as in visualizing how these shapes could be fitted together, like jigsaw pieces for example? Or does the

[1] Written by S/Sgt. Wayne S. Zimmerman.

latter task require some additional kinesthetic ability as well? These are examples of problems of fundamental interest to psychological science, and if they can be solved, the results will also be of considerable practical significance. (1)

Early Factorial Studies

The technique of factor analysis assumes the existence of separable abilities, and furnishes a comparatively objective tool for analyzing data in the light of this assumption. Perhaps it is from this objective technique that the most promising evidence is to come for the establishment of visualization as an independent mental ability or set of abilities. Both Kelley (2) and Thurstone (1) found evidence in early factor studies of the ability to visualize. Thurstone referred to a factor which he thought could be characterized justifiably as "visual and spatial imagery." He noted that it was probably the same factor that had appeared previously as a spatial factor in the studies of Kelley. For some reason, the term "spatial ability" rather than "visualization" was employed by these investigators and their followers. It is probable that substantial variance of both spatial and visualization factors was included.[3]

Visualization in Aviation Psychology

Perhaps the first mention of visualization as an important pilot ability occurred in the study of causes of failure of 1,000 cadets eliminated from pilot training. In a study of faculty-board proceedings, it was found that in 43 percent of the cases, deficiency in visualization of the flight course was mentioned as a reason for elimination. Visualization of the flight course was described as the ability to "get out of the cockpit and fly the plane with regard to the horizon and other reference points."

The first appearance of a factor that was later recognized as visualization was in an analysis of a battery of nonverbal reasoning tests at Psychological Research Unit No. 3.[4] On one axis, the following four tests had loadings of 0.50 or more: Spatial Visualization I, 0.58; Pattern Comprehension, 0.51; Mechanical Principles, 0.50; Spatial Visualization II, 0.50. By inspection, these tests seem to have only one element in common, that of the visual manipulation of images in solving the problems.

Following this analysis, some of Thurstone's data were reexamined, and an analysis was made of 19 of his tests based upon the published matrix of intercorrelations. A visualization factor was isolated and defined by the following tests: Punched Holes, 0.58; Form Board, 0.50; Lozenges B, 0.45; Surface Development, 0.40. Three subsequent analyses, including two of perceptual tests and one of mechanical tests, further supported the belief in the existence of a visualization factor. Mechani-

[3] In Thurstone's published analysis of 56 variables, an inspection of the final loadings reveals that further rotations of residual axis 12 with other identified axes would produce a promising visualization factor, with Punched Holes, Lozenges A, and Paper Form Board, in that order, appearing with the most significant loadings.

[4] See chapter 7 for the complete presentation of this analysis.

cal Principles, Map Distance, Mechanical Comprehension, and Shortest Path defined the factor in these analyses.

Tests that have been found to measure some phase of the ability to visualize are described and discussed in this chapter under the subheadings (1) Visual Manipulation and (2) Visual Completion.

VISUAL MANIPULATION TESTS

Tests described within this group have in common problems that appear to demand mental manipulation of a visual image or images. This type of visualizing calls for an ability to imagine the rotation of depicted objects, the folding or unfolding of flat patterns, the relative changes of position of objects in space, the motion of machinery, or the maneuvering of airplanes in space. In all tests in this group, the examinee's task is to record the final position or positions after a visualized movement or manipulation has taken place.

Under this heading one commercial test and eight experimental tests are discussed in the order in which they were chosen for study. They are: Pattern Comprehension, CP803A, 803B; Spatial Visualization II, CI203AX1, 203A; Spatial Visualization I, CI204AX1, 204AX2; Visualization of Maneuvers, CI657AX1, BX1, CX1, CX2; Formation Visualization, CP814A, 814AX2; Point Motion (Crawford-Bennett), Form B; Spatial Visualization III, CP108A; Position Visualization I, CP534A; and Position Visualization II, CP111A.

Pattern Comprehension, CP803A [4]

This test was adapted by the AAF from Thurstone's "Surface Development." Since it involved the visualized folding of flat patterns into three-dimensional objects, it might be expected to be a measure of manipulative visualization. At the time of its adoption, it was of interest primarily because surface-development tests had traditionally been included in mechanical-test batteries.

Description.—A flat pattern lay-out, showing the outline in solid lines and the edges along which the folds are to be made in dotted lines, is presented alongside of an isometric drawing of the three-dimensional object that would be formed by folding the pattern correctly. Certain edges on the folded object are numbered, while edges and fold lines on the flat pattern are lettered. The examinee's task is to match the numbers and the letters. In order that inside and outside surfaces will not be confused, two corresponding adjacent edges on the folded figure and on the flat pattern are marked, one with an X and the other with an O.

(1) *Internal characteristics.*—The directions contain one illustration of a flat pattern, the accompanying illustration of the three-dimensional figure, and two recorded but unscored sample questions. The test contains 7 patterns and 32 scored questions.

[4] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis and Lt. Linn Hutchinson.

(2) *Administration.*—All the necessary directions are contained in the booklet. An answer legend which shows five of the letters from the pattern listed as alternatives A, B, C, D, and E is provided with each item in the booklet. Fifteen minutes are allowed to complete the items in the booklet. The administration time is 5 minutes, making a total required testing period of approximately 20 minutes.

A sample item from form CP803A is shown in figure 12.1. Following are part of the directions:

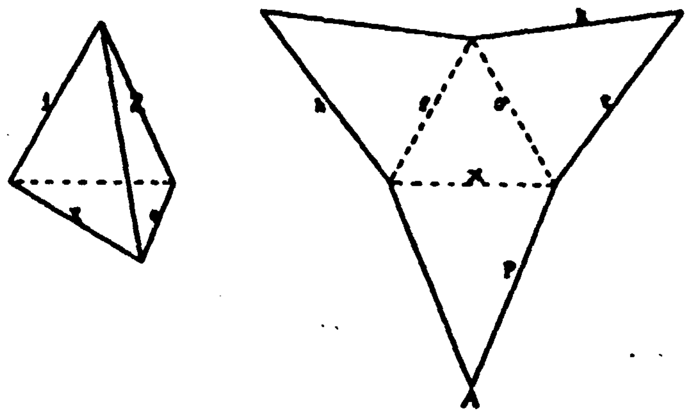


FIGURE 12.1
SAMPLE PROBLEM OF PATTERN COMPREHENSION, CP803A

This is a test to see how quickly and accurately you can understand the relationship between a pattern drawing and the object it represents.

In the example are two drawings, one showing a solid figure and the other showing a plane figure. If the solid at the left is placed on the figure at the right, the latter can be folded perfectly around the solid. The figure at the right may therefore be called a pattern of the solid at the left. In the pattern the area bounded by dotted lines corresponds to the base of the solid. Two of the edges of the solid and the two corresponding dotted lines in the pattern have been marked X and O. Using these two edges for reference, select the edges of the pattern which correspond to each of the numbered edges of the solid and mark the answers opposite the problem numbers on your answer sheet.

1. Corresponds to (A) h; (B) p; (C) f; (D) t; (E) k.
2. Corresponds to (A) t; (B) f; (C) h; (D) k; (E) p.

The correct answers are 1, A; 2, D.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data given below are for samples tested at Psychological Research Unit No. 3 in March and April 1943.

(1) *Distribution statistics.*—A sample of 229 unclassified aviation students yielded a mean score of 14.9 and a standard deviation of 8.3. The distribution curve was approximately symmetrical and considerably flatter than normal.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.42, a standard deviation of the phi distribution of 0.06, and a range of values from 0.28

to 0.58. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 375 unclassified aviation students.

(3) *Difficulty.*—Based upon item analysis of the responses of 800 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.50, corrected for chance, with a range of 0.13 to 0.79 and a standard deviation of 0.16.

(4) *Factorial composition.*—The most significant loadings are in the visualization (0.50), general-reasoning (0.33), perceptual-speed (0.24), and reasoning II (0.24) factors. The communality is 0.55. For a full picture of the factorial composition of this test, see appendix B.

(5) *Test validity.*—Validation results based on two samples are given in table 12.1.

TABLE 12.1.—*Validity data for Pattern Comprehension based upon graduation-elimination of pilots in primary training*

Form	N_t	Class	p_g	M_g	M_e	SD_t	r_{bis}
CP803A	211	44A	0.86	15.34	12.40	8.28	0.19
CP803AX2[1]	652	43K, 44A	.86	24.36	20.97	11.22	.16

[1] Items identical with those in CP803A.

Variations of the test.—The first form of Pattern Comprehension, CP803AX1, contained, in addition to the sample pattern and questions, 11 patterns and 72 scored items. The AX2 form was made up of the items from the original form which proved to have the highest internal consistencies. The directions remained unchanged. Seven patterns and 32 questions were selected. The time limit was reduced from 30 to 15 minutes. Both of these test forms were mimeographed. The A form was printed rather than mimeographed. Except for revised directions, it is identical with the shortened AX2 test.

In form CP803B[8] several changes were incorporated. In order that the X and O designations on the patterns, which are necessary to define whether a diagram represents an inside or an outside pattern, might be eliminated, all diagrams are drawn as inside patterns. The necessity for including a different answer legend for each question was removed by adapting the answers to the 15-place IBM answer sheet. A proportionately smaller number of questions per pattern is asked, 30 questions for 10 patterns being included for the entire test. The test is divided into two equivalent parts, separately timed. Only two patterns were retained from the preceding forms. The newly constructed diagrams are all asymmetrical, in contrast with the symmetrical construction of the original patterns, and they also have fewer sides. Both of these changes were made in the effort to eliminate some of the reasoning content. It was hypothesized that the symmetrical diagrams afforded more opportunity for the examinee to derive answers by noting reverse relationships,

[8] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Lloyd G. Humphreys, Sgt. Fred H. Meise, and S/Sgt. Wayne S. Zimmerman.

which could be reasoned through to a solution. No data are available on this form.

Evaluation.—Pattern Comprehension, originally selected for study because it was thought to be related to mechanical ability, proved to be a fairly good measure of manipulatory visualization. The high correlation of Mechanical Principles and Pattern Comprehension ($r=0.42$) is largely due to the saturations with this factor. A source of dissatisfaction has been the high loading of Pattern Comprehension with general reasoning. It is hoped that form B will prove to be somewhat less tainted with that factor.

Pattern Comprehension shows moderate pilot validity, the weighted validity coefficient being 0.16 for a total sample of 1,081 pilots on forms A and AX2. From its known factor loadings and their validity for the pilot, one would expect a validity of 0.14. The predicted validity of the AX1 form is 0.10; the obtained weighted validity is 0.09 for a total of 525 cases.

It is interesting to note that the loading in the mechanical-experience factor is only 0.06, as based on a weighted mean of the loadings found in two analyses. Whatever virtue this test may have for selection for mechanical tasks, therefore, probably would arise from its perceptual-speed and visualization loadings. If so, there are much better tests for that type of selection.

Spatial Visualization II, CI203A [a]

This test was adapted from the Verbal Cubes test which was prepared before the war by Col. J. P. Guilford. It was selected for study because it promised to measure nonverbal reasoning. A battery of nonverbal reasoning tests was being assembled for intercorrelational and factor analysis (see ch. 7). Verbal Cubes was originally designed to be a measure of the ability to manipulate mental images. This concept was only an incidental consideration, however, at the time the revision was begun by Psychological Research Unit No. 3.

Description.—For each group of items, the examinee reads a verbal description of a solid block of wood, its sides painted different colors, which is cut into smaller blocks. The examinee's task is to visualize these cutting operations so that he can answer questions about the resulting numbers of blocks of given size and color.

(1) *Internal characteristics*.—The directions contain one verbal problem description accompanied by two recorded but unscored sample questions. Parts I and II each contain 6 descriptions and 22 scored items. The problems increase in difficulty toward the end of each part.

(2) *Administration*.—On each page of the booklet is an answer legend for converting numbers into letters to correspond to the letters

[a] Developed at Psychological Research Unit No. 3. Chief contributors: S/Sgt. Jacob G. Etkin, Lt. David H. Jenkins, and S/Sgt. Wayne S. Zimmerman.

on the 15-place answer sheet. Thirteen minutes are allowed for each part.

Following are sample items and accompanying explanation taken from the directions from the test:

The ends of a block of wood 1" x 1" x 3" are painted black, and the block is then cut into 1-inch cubes

Answer Legend

A. 1
B. 2
C. 3
D. 4
E. 6

1. How many 1-inch cubes are there?
2. How many 1-inch cubes have only one side painted black?

If you pictured the pieces of wood correctly, you should have marked "C" for item 1 and "B" for item 2, since the 3-inch piece can be cut into three 1-inch cubes but only the two end cubes would have painted sides.

(3) *Scoring.*—The scoring formula is $R-W/5+20$.

Statistical results.—The data given below are for examinees tested at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Distribution statistics for two overlapping samples are given in table 12.2. The distribution curves are moderately positively skewed and somewhat flatter than normal.

TABLE 12.2.—*Distribution constants for Spatial Visualization II, C1203A, based upon samples of classified pilots in class 44G*

N	M	SD
2,115[1]	34.7	10.2
1,390[1]	36.4	10.1

[1] Overlapping samples.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.50, a standard deviation of 0.11, and a range of values from 0.19 to 0.67. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score on form AX1 of a group of 450 unclassified aviation students, tested in April and May 1943.

(3) *Reliability coefficient.*—By the alternate-forms method, (part I-part II), a reliability coefficient of 0.84, corrected for length, was obtained. This figure is based on a sample of 487 classified pilots in class 44G, tested in January 1944.

(4) *Correlation between rights and wrongs.*—For a sample of 500 classified students (tested in 1944; specific dates of testing not reported) the correlation between rights and wrongs was −0.67. For 594 navigators tested in May 1944, the correlation was −0.66.

(5) *Difficulty.*—Based upon item analysis of the responses of the above-mentioned sample of 450 unclassified aviation students, form AX1

TABLE 12.3.—*Validity data for Spatial Visualization II, CI203AX1 and CI203A*

Group	Form	Scoring formula	Criterion	N_t	p_s	M_p	M_e	SD_t	r_{bis}	$_c r_{bis}$
Pilots in primary training[1]	CI203AX1	R−W/5	Graduation-elimination	774	0.93	16.16	12.30	9.60	0.19	[2]0.22
Pilots in primary training[3]	CI203A	R−W/5+20	"	2,115	.89	34.90	32.75	10.19	.11	[4].15
Navigators[5]	CI203A	R	"	594	.89	23.00	20.17	8.07	.18	[6].35
Navigators[5]	CI203A	W	"	594	.89	14.29	16.20	7.01	−.14	[6]−.27
Navigators[5]	CI203A	R−W/5	"	594	.89	20.14	16.93	9.06	.19	[6].36
Navigators[7]	CI203AX1	R−W/5+20	Flight mission grades	200	....	.....	.....	.....	[8].18	.....
Navigators[7]	CI203AX1	R−W/5+20	Ground mission grades	200	....	.....	.....	.....	[8].37	.....
Navigators[7]	CI203AX1	R−W/5+20	Weighted total grades	200	....	.....	.....	.....	[8].34	.....

[1] In class 44C and 44D.
[2] Assumed unrestricted stanine standard deviation not reported.
[3] In class 44G.
[4] Assuming an unrestricted stanine standard deviation of 1.87.
[5] Same sample, tested in May 1944.
[6] Assuming an unrestricted stanine standard deviation of 2.00.
[7] Same sample. Group not identified. Data reported by Psychological Research Unit No. 3 in October 1943.
[8] Product-moment correlation.

yielded a mean proportion of correct responses of 0.51, corrected for chance, with a range from 0.11 to 0.86 and a standard deviation of 0.19.

(6) *Factorial composition.*—The most significant loadings are in general-reasoning (0.44), visualization (0.42), reasoning III (0.36), and reasoning II (0.35) factors. The communality is 0.75. For a full picture of the factorial composition of this test see appendix B.

(7) *Test validity.*—Validation results based on several samples are given in table 12.3.

(8) *Item validity.*—Validation of items revealed a mean phi of 0.03 based upon the responses of 600 graduates and 127 eliminees from primary pilot training originally tested in December 1943 and January 1944. The standard deviation of phi values was 0.05, and the range was from −0.08 to 0.15.

Evaluation.—Spatial Visualization II is a fairly good measure of manipulatory visualization; but, like Pattern Comprehension, it is so complicated by reasoning factors that it is of little value in predicting pilot success. Twenty-two percent of its total variance is attributable to visualization, 22 percent to reasoning I, 14 percent to reasoning II, and 15 percent to reasoning III. The average obtained validity of 0.17 exceeds slightly that to be expected from its loadings with known factors (0.12). This fact, taken together with the difference between the reliability and the communality of the tests, suggests some unknown source of validity. The very large proportion of reasoning variance, however, renders this test unfit for use in pilot selection.

Because of its combination of factors, it should have a validity of at least 0.34 for navigation training. The limited data in table 12.3 almost exactly fulfill this expectation. Because it is a complex test, however, it should not be included in a battery when the important factors are already covered by purer tests.

Variations of the test.—CI203AX1 and CI203A [7] are identical, except for certain changes in the directions. In the A form, an attempt was made to add face validity by pointing out in the directions how the task presented in the test is related to flying problems.

Spatial Visualization I, CI204AX1 [8]

When the antecedent of this test (Paper Folding) was devised by Col. J. P. Guilford before the war, it was set up to be a measure of visualization, but because of high intercorrelations with reasoning tests, it was adapted by the AAF for the purpose of studying nonverbal reasoning ability.[9]

Description.—For each item, two or three illustrations show step by step how a square, circular, or triangular piece of paper is folded and

[7] Directions written by Sgt. Nathan Kravetz.
[8] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Frank Dudek and S/Sgt. Wayne S. Zimmerman.
[9] See chapter 7.

finally cut. Blackened areas represent portions that are cut out after the final fold has been made.

The examinee's problem is to determine how the piece of paper will look when it is unfolded. To the right of the illustration of the folds are five representations of unfolded figures. One of the five unfolded figures correctly shows all the creases made by folding and all the holes made by cutting.

(1) *Internal characteristics.*—The directions contain one recorded but unscored practice item. Each part contains 30 scored items. Items in part I are made up of square pieces of paper and the items in part II are made up of circular and triangular pieces.

(2) *Administration.*—The directions consume approximately 5 minutes, while 20 minutes are recommended for completion of the items in part I and 19 minutes for the items in part II, making a total testing time of approximately 44 minutes.

In figure 12.2 is shown the sample item used in the test.

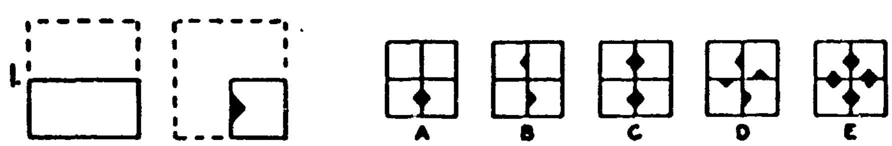


FIGURE 12.2
SAMPLE PROBLEM OF SPATIAL VISUALIZATION I, CI204AXI

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data given below are based upon samples tested at Psychological Research Unit No. 3, except where noted.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 12.4. The distribution curves are moderately negatively skewed and considerably flatter than normal.

TABLE 12.4.—*Distribution constants for Spatial Visualization I, CI204AX1*

Group	N	M	SD
Unclassified aviation students[1]	242	35.2	10.0
Unclassified aviation students[2]	204	33.4	9.7
Classified pilots[3]	196	33.4	9.6

[1] Tested in May 1943.
[2] Tested in March and April 1943
[3] In class 44A.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.35, a standard deviation of the phi distribution of 0.14, and a range of values from −0.12 to 0.64. These statistics are based upon the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 450 unclassified aviation students tested in August 1943.

(3) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.84, corrected for length, was obtained. This figure is based on a sample of 203 unclassified aviation students tested in April 1943.

(4) *Correlation between rights and wrongs.*—For a sample of 735 navigators tested in February 1944 at Ellington Field and at Psychological Research Unit No. 3, the correlation between rights and wrongs was −0.57.

(5) *Difficulty.*—Based upon item analysis of the responses of 450 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.67, corrected for chance, with a range from 0.04 to 0.97 and a standard deviation of 0.34.

(6) *Factorial composition.*—The most significant loadings are in the visualization (0.56), reasoning I (0.34), reasoning II (0.32), space I (0.24), and reasoning III (0.21) factors. The communality is 0.71. For a full picture of the factorial composition of this test see appendix B.

(7) *Test validity.*—Validation results based on several samples are given in table 12.5.

TABLE 12.5.—*Validity data for Spatial Visualization I, CI204, using the graduation-elimination criterion*

Group	Form	Scoring formula	N_t	r_{tt}	M_s	M_f	SD_t	r_{bis}	$_c r_{bis}$[1]
Pilots in primary training[2]	CI204AX1	R − W/4	196	0.86	34.06	29.34	9.64	0.26	....
Pilots in primary training[3]	CI204AX2	R − W/4	1,249	.91	19.46	18.46	8.01	.06	0.13
Navigators[4]	CI204AX2	Rights	735	.88	26.16	22.74	7.05	.25	.44
Navigators[4]	CI204AX2	Wrongs	735	.88	7.36	10.00	6.01	−.23	−.39
Navigators[4]	CI204AX2	R − W/4	735	.88	24.32	20.24	8.00	.27	.45

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44A.
[3] In classes 44F, 44G, and 44H.
[4] Tested at Ellington Field and Psychological Research Unit No. 3 in February 1944.

Variations.—The AX2[10] form is made up of 40 items selected on the basis of internal-consistency phis from the 60-item AX1 form. Items are arranged in order of increasing difficulty, which was determined from the item analysis. Its leading factors and their loadings were found to be visualization (0.53), general reasoning (0.39), integration III (0.34), and the verbal factor (0.26). The communality was 0.69. There is no ready explanation for the discrepancies between the two forms in secondary factors.

Evaluation.—Spatial Visualization I proved to be one of the best measures of manipulatory visualization, although significant loadings on three reasoning factors and space I indicate a complex factorial pattern. Forty-three percent of the total variance is attributable to the visualization factor, 16 percent to general reasoning, and 14 percent to reasoning II. The pilot validity to be expected from its factorial picture is 0.19, and its obtained mean validity is 0.15. The two validities probably agree

[10] Developed at Psychological Research Unit No. 3. Chief contributor: Jean R. Lyons.

within the limits of error. The expected validity for the AX2 form is 0.15 for pilots, and the obtained validity is 0.12. The obtained validity for navigator selection is quite high (0.45). Since the reasoning I and visualization factors account for only half of this validity, and probably less, the test has something unique and substantial to offer in this connection.

This test suggested the construction of the promising Spatial Visualization III test, an orally administered test of the ability to solve paper-folding problems.

Visualization of Maneuvers, CI657AX1 [11]

This test was constructed as a possible measure of both visualization and integration.

Description.—Each item begins with the presentation of an airplane in a given attitude shown in a photograph. Then three simple maneuvers are stated; such as, turn right 90°, nose up 45°, roll left 90°. Imagining himself as the pilot of the plane, the examinee must visualize these maneuvers in sequence, beginning from the pictured original position. Then he must select the correct final position of the plane from five alternative positions shown pictorially. Thus, the examinee must keep in mind the changed position of the plane after the first maneuver and from this position visualize the second maneuver. Again, he must hold the new position in mind and from this second position visualize the third maneuver. All maneuvers must be visualized from the pilot's position in the cockpit, i. e., turn right means to the pilot's right, regardless of the plane's position on the page.

(1) *Internal characteristics.*—The directions contain two recorded but unscored sample items. Part I contains 28 scored items, and part II contains 30.

(2) *Administration.*—Twenty-five minutes are allowed for each part of the test. Directions and sample items consume about 10 minutes, making a total testing time of 60 minutes.

One sample item is shown in figure 12.3. Following are parts of the directions:

> This is a test of your ability to visualize airplane maneuvers. In each problem the pilot of a plane will take it through three maneuvers. On the left is shown the starting position of the plane and on the right are shown five positions, one of which is the final position of the plane after the maneuvers have been executed.
>
> The completion of the third maneuver puts the plane in the position shown in picture C. C is, therefore, the correct answer.

(3) *Scoring.*—The scoring formula used is R−W/4.

Statistical results.—Except where specifically noted to the contrary, the following data are based upon samples tested at Psychological Research Unit No. 3.

[11] Developed at Psychological Research Unit No. 3. Chief contributor: S/Sgt. Wayne S. Zimmerman.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 12.6. The distribution curves are approximately symmetrical and considerably flatter than normal.

TABLE 12.6.—*Distribution constants for Visualization of Maneuvers based upon samples of classified pilots*

Form	Number of items	N	M	SD
AX1	58	[1]1,162	27.4	14.3
BX1[2]	58	[3]1,222	35.2	13.5
CX1[2]	98	[4]1,390	56.6	16.9

[1] In class 44F.
[2] For descriptions of these forms see page —.
[3] In class 44H.
[4] In class 44G.

(2) *Internal consistency.*—The degree of homogeneity of the items in form AX1 is indicated by a mean internal-consistency phi of 0.52, a standard deviation of the phi distribution of 0.10, and a range of values

TABLE 12.7.—*Alternate-forms (part I v. part II) reliability coefficients for Visualization of Maneuvers, CI657*

Group	N	Form	r'_{tt}	r_{tt}
Classified pilots[1]	1,619	AX1	[2]0.81	0.89
Classified pilots[3]	523	BX1	[2].76	.86
Unclassified aviation students[4]	193	CX1	[5].82	.90
Unclassified aviation students[4]	525	CX1	[2].85	.92
Unclassified aviation students[4] and Airplane Mechanics	448	CX1	[2].84	.91
Classified pilots[6]	504	CX1	[2].77	.87

[1] In class 44F.
[2] Part II administered immediately after part I.
[3] In class 44H.
[4] Tested at Medical and Psychological Examining Units Nos. 6 and 8 in April 1945.
[5] Part II administered approximately four hours after part I.
[6] In class 44G.

from 0.26 to 0.80. These statistics are based upon analysis of the responses of the highest 25 percent and the lowest 25 percent in total score of a group of 800 classified pilots in class 44F.

(3) *Reliability coefficient.*—The three forms yielded the alternate-forms estimates of reliability given in table 12.7.

(4) *Correlation between rights and wrongs.*—Data are presented in table 12.8.

TABLE 12.8.—*Correlations between rights and wrongs for Visualization of Maneuvers CI657*

Group	Form	N	r_{rw}
Classified students[1]	AX1	500	−0.79
Classified students[1]	BX1	500	−.73
Classified pilots[2]	CX1	642	−.35
Navigators[3]	CX1	1,713	−.39

[1] Tested in 1944; specific testing dates not reported.
[2] Tested May 9 to Aug. 12, 1944.
[3] Tested in June 1944 at Psychological Research Unit No. 1, in April 1944 at Psychological Research Units Nos. 2 and 3.

Table 12.9.—*Validation data for Visualization of Maneuvers, graduation-elimination criterion*

Group	Form	Scoring formula	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
Pilots in primary training[a]	AX1	R−W/4	1,162	0.94	27.77	20.98	14.25	0.23	0.30
Pilots through basic training[b]	AX1	R−W/4	1,145	.90	27.98	21.81	14.25	.22	.31
Pilots in primary training[c]	BX1	R−W/4	1,122	.87	36.36	30.32	13.59	.24	.29
Pilots in primary training[d]	CX1	Rights	642	.77	52.57	46.08	16.21	.23	.34
Pilots in primary training[d]	CX1	Wrongs	642	.77	23.79	23.66	9.41	.01	−.05
Pilots in primary training[d]	CX1	R−W/4	1,390	.91	57.27	49.90	16.88	.22	.27
Navigators[e]	CX1	Rights	1,713	.91	55.80	50.98	15.55	.16	.34
Navigators[e]	CX1	Wrongs	1,713	.91	22.41	24.97	9.43	−.14	−.20
Navigators[e]	CX1	R−W/4	1,713	.91	50.20	44.74	16.60	.17	.35

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[a] In class 44F.
[b] In classes 44H and 44I.
[c] Tested May 9 to Aug. 12, 1944.
[d] In class 44G. Tested January 1944.
[e] Tested in June 1944 at Psychological Research Unit No. 1, and in April 1944 at Psychological Research Units Nos. 2 and 3.

(5) *Difficulty.*—Based upon item analysis of the responses of the above-mentioned sample of 800 classified pilots, the AX1 form of the test yielded a mean proportion of correct responses of 0.46, corrected for chance, with a range from 0.19 to 0.84 and a standard deviation of 0.14.

(6) *Test validity.*—Validation results for the three forms of Visualization of Maneuvers are given in table 12.9.

(7) *Item validity.*—Validation of items on two forms of this test disclosed the results recorded in table 12.10.

TABLE 12.10.—*Validity of items of Visualization of Maneuvers based upon samples of pilots in primary training, graduation-elimination criterion*

Form	N_t	r_e	$M\phi$	$SD\phi$	Range of ϕ	
					Low	High
AX1	[1]548	0.88	0.10	0.08	−0.09	0.28
AX1	[2]675	.89	.04	.07	−.07	.22
CX1	[3]727	.83	.04	.07	−.08	.15

[1] In class 44H.
[2] In class 44F.
[3] In class 44G.

Variations.—Form BX1,[12] CXI,[13] and CX2 differ from form AX1 only in the number of maneuvers called for between the initial and final positions of the airplanes. Form BX1 presents two maneuvers; both forms CX1 and CX2 present one maneuver. These other forms were developed in an effort to lessen the difficulty of the items and consequently to lay more dependence upon speed. It was hypothesized that good pilots excel in acts that are undeliberated rather than reasoned and that speeded tests, therefore, are likely to show more valid results. Forms AX1 and BX1 are of the same length, each containing 58 scored items divided into 2 parts. Form CX1 contains 98 scored items divided into 2 parts, while CX2 contains the first 48 items from form CX1 divided into 2 parts.

Forms AX1 and BX1 correlate more highly with Mechanical Principles and less highly with Complex Coordination than form C, a fact that supports an original hypothesis advanced during the construction of the test that the more complex form would show more visualization, the speeded form, more spatial content. This evidence needs the further support of factor analysis.

Evaluation.—The Visualization of Maneuvers tests proved to be one of the most valid types of printed tests for pilots in or out of the classification battery. It would be an excellent selection test for either pilots or navigators, but it would not make a good classification test in which a discriminating function is desired.

[12] Developed at Psychological Research Unit No. 3. Chief contributor: S/Sgt. Wayne S. Zimmerman.
[13] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Stuart W. Cook and S/Sgt. Wayne S. Zimmerman.

No factorial data are available at the time of this writing, but intercorrelations suggest that the AX1 form is comparatively complex and would, therefore, be used more appropriately in an omnibus test like the AAF Qualifying Examination. The level of validity for both pilot and navigator also assures us that the test is factorially complex and combines factors that are strongly valid for both specialties. For pilot validity, both space and visualization must surely be present in large amounts. For navigator validity some reasoning variance must also be present.

Formation Visualization, CP814A [14]

This test was developed in an effort to measure the examinee's ability to visualize in three dimensions. If the views of airplanes in formation are shown from two directions at right angles to each other, it is possible by visualization to determine the appearance of the same formation from the third orthogonal direction. Since this type of item presents a rather difficult visual manipulative problem, only a limited number of airplanes can be presented in each formation. It is known that overly difficult visualization items are likely to be reasoned through to a solution. In order to keep the difficulty level as low as practicable, only two or three airplanes were included in any single formation. The use of airplanes as the objects to be visualized adds face validity to the test.

Description.—Each item shows in silhouette a top view and a side view of a formation of either two or three airplanes. The examinee's problem is to visualize the appearance of the same formation from the front view. Four alternative front views are presented with each item, one of which is correct.

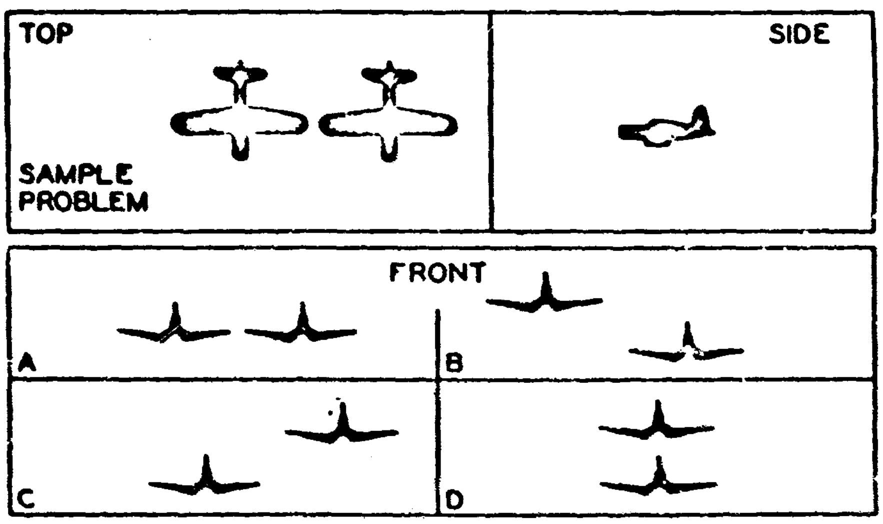


FIGURE 12.4
SAMPLE PROBLEM OF FORMATION VISUALIZATION, CP814A

[14] Developed at Psychological Research Unit No. 3. Chief contributors: Cpl. Albert A. Canfield and Maj. Robert L. Thorndike.

(1) *Internal characteristics.*—The directions contain one recorded, but unscored, sample item. The test contains 49 scored items. The first 19 items are composed of formations of 2 airplanes, while all of the remaining items are made up of formations of 3 airplanes.

(2) *Administration.*—Administration of the directions takes 2 minutes, and 15 minutes are allowed to complete the test items, making a total testing time of 17 minutes.

The sample problem from form CP814A is shown in figure 12.4. Following are parts of the directions:

This is a test of your ability to visualize plane formations. In each problem you will see two views of a formation of either two or three planes. One view will show the formation as seen from above; the other, the formation as seen from the side. Your task is to visualize how this formation would appear if it were seen from the front.

Below each formation there are four front-view diagrams, A, B, C, and D, only one of which correctly represents the formation as it would appear from the front. These diagrams are not drawn to scale. Remember that only one of the four diagrams in each item represents the front view of the formation.

Which diagram correctly represents the front view of the formation in the sample problem?

As only one plane can be seen in the side view, the other plane must be concealed directly behind it. Therefore, the two planes in this formation should be visualized as flying side by side at the same altitude.

The correct front view of this formation is shown by diagram A.

(3) *Scoring.*—The scoring formula used is $R-W/3+20$.

Statistical results.—The data given below are for samples tested at Psychological Research Unit No. 3 in September and October 1944.

(1) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.36, a standard deviation of the phi distribution of 0.12, and a range of values from 0.13 to 0.66. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 1,500 unclassified aviation students.

(2) *Difficulty.*—Based upon the responses of 750 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.57, corrected for chance, with a range from 0.14 to 0.94, and a standard deviation of 0.23.

Evaluation.—No further data are available upon which conclusions can be based regarding the nature of this test. Both the internal consistency and difficulty levels are satisfactory.

Crawford-Bennett Point Motion Test [18]

Imagining the motion of machinery and following these motions mentally has been described as a demonstration of the ability to visualize.

Description.—Each item presents an assembly drawing of certain parts of a machine. The path that a single point on the machine will

[18] Published by the Psychological Corp., New York, N. Y.

follow when the mechanism is set in motion is indicated. A second point in another section of the machine is also marked, but the path for this point is not shown. It is the examinee's problem to determine the exact path that the second point will follow when the mechanism is in operation. The correct path must be selected from four illustrated choices.

(1) *Internal characteristics.*—The test contains 1 unrecorded and unscored sample item in the directions and 30 scored items in the body of the test.

(2) *Administration.*—When this test was first administered to prospective air-crew members, 25 minutes were allowed to complete the items. The testing period was found to be unnecessarily long and was later reduced to 15 minutes.

The sample item is shown in figure 12.5. Following are excerpts from the directions:

How will point X move when point P moves as shown by the arrows? Choose your answer from A, B, C, or D.

B is the correct answer, since the curve B best describes the motion of point X.

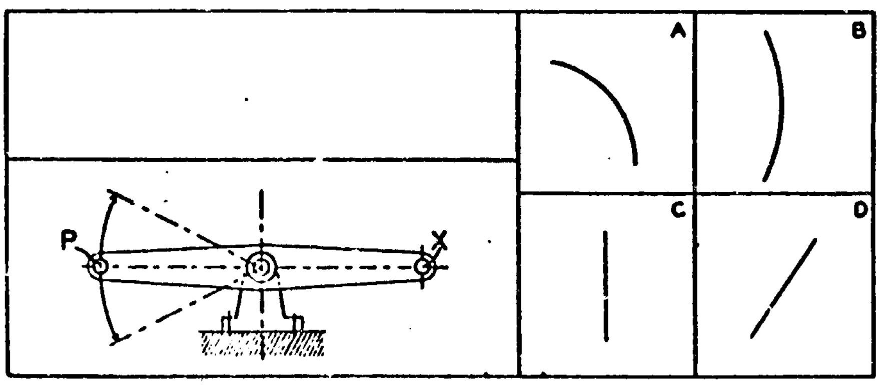


FIGURE 12.5
SAMPLE PROBLEM OF CRAWFORD-BENNETT POINT MOTION, FORM B

(3) *Scoring.*—The scoring formula is R−W/3.

Statistical results.—This test was administered at Psychological Research Unit No. 3.

(1) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.38, a standard deviation of 0.18, and a range from 0.02 to 0.69. These statistics are based upon analysis of the responses of highest 27 percent and the lowest 27 percent of 740 unclassified aviation students tested in June 1944.

(2) *Difficulty.*—Based upon the responses of the above-mentioned sample of 740 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.48, corrected for chance, with a range from 0.03 to 0.94 and a standard deviation of 0.26.

(3) *Test validity.*—Validation results are given in table 12.11.

TABLE 12.11.—*Validity data for the Crawford-Bennett Point Motion Test, form B, based upon elimination from primary training* [N_t=973; p_s=0.80]

Score	M_p	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
Rights[2]	17.12	16.62	3.91	0.07	0.20
Wrongs[2]	12.43	13.16	4.06	−.10	−.22
R−W/3	12.98	12.23	5.19	.08	.20

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] For this sample, the correlation between rights and wrongs is −0.93.

Evaluation.—A validity of −0.22 computed on the wrong scores and of 0.20 computed on the right scores indicates that this test does have some pilot validity, as was predicted. The validity was not high enough, however, to add significantly to the prediction efficiency of the classification battery in view of its high correlation with the pilot stanine.

Spatial Visualization III, CP535 [16]

Following a period of concentrated factor study, a program was outlined for the development of tests to measure the known factors of intellectual ability in as pure a fashion as possible. This test is one of a group of tests designed in an attempt to secure a pure measure of the visualization factor.

The paper-folding test, Spatial Visualization I, proved to be one of the best available measures of the factor, although it was more highly saturated with reasoning than was desired. It was hypothesized that (1) the reasoning content of Spatial Visualization I is primarily due to the opportunity afforded in the pictorial presentation to note relationships and to derive systems for obtaining answers without depending upon visualizing powers, and (2) a verbal presentation would reduce the opportunity to solve the problems by any method other than visualization.

Description.—From an orally delivered description, each item requires the subject to visualize the folding of a square sheet of paper into various shapes. The final correct shape must be selected from drawings presented in the test booklet.

(1) *Administration.*—The oral descriptions are presented by means of phonograph records. The examinees listen to the recorded test items with their booklets closed. When the description of the paper folding is completed, the examinee is instructed to open his booklet. He is then told the number of the item in the booklet, after which the correct choice can be found among five alternative illustrations. Ten seconds are allowed for locating and recording each answer. Then the examinee is instructed to close his test booklet before the next problem is presented.

The five alternative illustrations for the sample item are shown in figure 12.6. Following are excerpts from the directions:

[16] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Gerald H. Shirley and S/Sgt. Wayne S. Zimmerman.

This is a test to see how well you can visualize. In every problem you are to imagine folding a square sheet of paper into various shapes. Since the directions cannot be repeated, you must listen very closely and follow each move as it is given.

Listen to sample problem one.

Imagine a square sheet of paper. Now imagine folding it in the middle from top to bottom. Now fold the upper left corner to the middle of the lower edge.

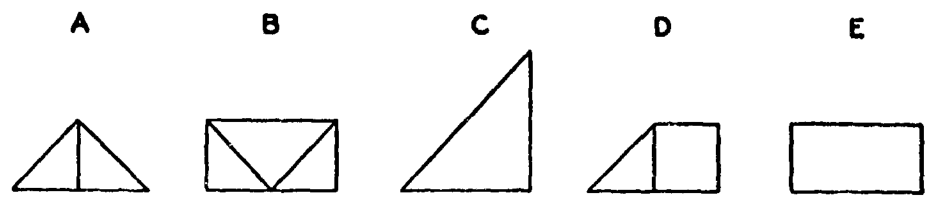


FIGURE 12.6
ALTERNATIVE ANSWERS OF THE SAMPLE PROBLEM, SPATIAL VISUALIZATION III, CP108A

Turn to page two, number eight in your answer booklet; page two, number eight.

Look at the five alternatives listed. Which alternative looks like the paper after it has been folded? Choice D is correct. Blacken in space D after number one on your answer sheet.

Close your booklet.

(2) *Scoring.*—The scoring formula is $R-W/4$.

Evaluation.—No data are available.

Position Visualization, CP534A [17]

This test is one of a group of tests designed for the specific purpose of obtaining a pure measure of manipulatory visualization. It was hypothesized that visual aids, such as drawings or illustrations of the objects to be visually manipulated, reduce the amount of visualizing required to solve the problems. Several tests were designed, therefore, with items presented in simple verbal terms so that the objects to be manipulated would have to be visualized without the help of visual cues. Due to the comparative ease with which its various positions could be described verbally, the United States flag was selected as the object to be visually manipulated.

Description.—Each item requires that a flag be visualized in a certain position. From the initial position the flag is to be both rotated and turned over according to specified directions. The examinee must visualize these manipulations and record the final position of the flag in terms of the location of the stars and the direction of the stripes.

(1) *Internal characteristics.*—The directions contain three recorded, but unscored, sample items. Part I contains 27 scored items, and part II contains 25 scored items. Items are printed verbally and are presented in tabular form.

[17] Developed at Psychological Research Unit No. 3. Chief contributors: S/Sgt. Benjamin Fruchter and Lt. John W. Howe, Jr.

(2) *Administration.*—Two minutes and 45 seconds are allowed to work the sample items. Twelve minutes are allowed for part I, and 10 minutes for part II. Reading directions takes about 5 minutes, making approximately 30 mintues total testing time.

Following are parts of the directions:

This is a test of your ability to imagine an object as it is moved from one position to another.

The object to be imagined is the American flag. At the start of each problem the position of the stars and stripes will be given; for example, stripes-horizontal; stars-upper left. Then you will be instructed to imagine moving the flag in certain definite ways, and in your answer describe the final position of the stars and stripes.

The flag will be moved only in the following ways:

The flag will be turned over the long way in some items and the short way in others. To turn over the flag simply means to switch the surfaces, as when you address an envelope and turn it over to seal it. Long way means the flag should be turned over lengthwise, as in illustration A. (See fig. 12.7.) Short way means it should be turned over crosswise, as in illustration B.

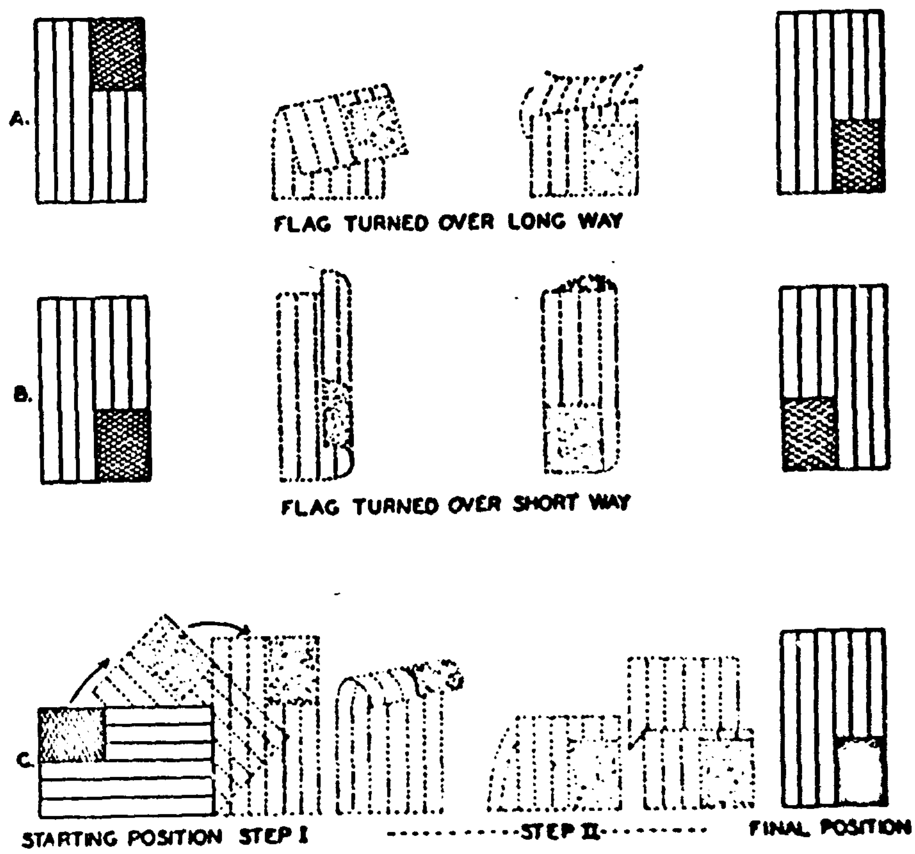


FIGURE 12.7
ILLUSTRATIONS OF FLAG TURNING AND FLAG ROTATING IN THE INSTRUCTIONS OF POSITION VISUALIZATION I, CP534A

The flag will be rotated 90° to the right in some items and 90° to the left in others. To the right means clockwise; to the left means counterclockwise. 90° means one-quarter turn.

The first practice problem is reproduced below:

Item	Starting position		Step I	Step II
	Stripes	Stars		
1.	Horizontal	Upper left	Rotate 90° to the right.	Turn over long way.

See the step-by-step solution to Practice Item 1 illustrated in figure 12.7.

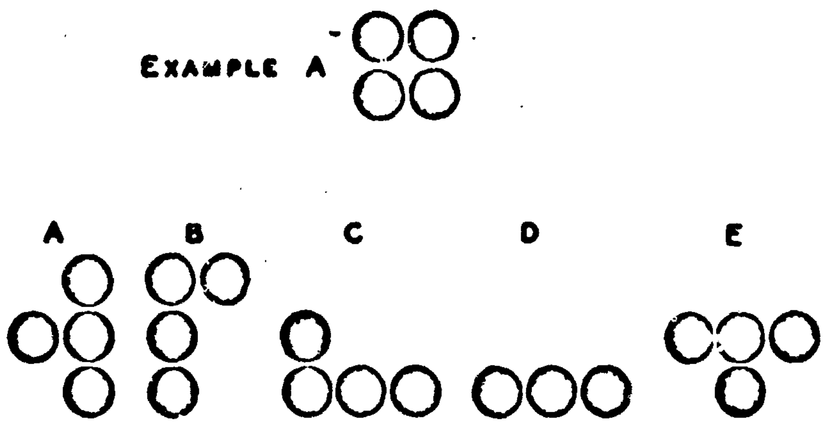


FIGURE 12.8
ALTERNATIVE ANSWERS FOR THE SAMPLE ITEMS, POSITION VISUALIZATION II, CPIIIA

To describe the position of the stars and stripes on your separate answer sheet, blacken the space under 'V' opposite question 1, because in the final position the stars and stripes are vertical. Blacken the space under 'LR' because in the final position the stars are in the lower-right corner. Every answer will have two parts; one to show the vertical or horizontal position of the stripes, and one to show the position of the stars.

Statistical results.—None are available.

Position Visualization II, CP111A[10]

This is another test designed for the express purpose of obtaining a pure measure of manipulatory visualization.

Description.—Each item requires the subject to visualize, in response to an orally delivered description, four objects (disks) forming a square. Then certain disks are to be imagined moved to different positions. The final pattern must be visualized, so that it can be correctly selected from patterns presented in the test booklet.

[10] Developed at Psychological Research Unit No. 3. Chief contributors: Cpl. Allan L. Berne and T/Sgt. Gerald M. Shirley.

(1) *Internal characteristics.*—The directions contain one unrecorded item and one recorded, but unscored, sample item. Parts I and II each contain 15 scored items.

(2) *Administration.*—Some of the directions and all of the problems are presented orally by phonograph record. The test booklet is closed during the reading of each problem. Directions to open the booklet and to look for the correct pattern following a specified number on a certain page follow the description of each problem. Directions to close the booklet precede each new problem. Administrative directions take approximately 5 minutes. Five minutes and 50 seconds are allowed for each part, making a total testing time of 17 minutes.

Following are parts of the directions:

In this test you are to imagine four disks forming a square.

First, imagine moving the lower left disks above the upper right disks. Now, imagine moving the bottom disk to the left of the left-most disk.

Next, look at the five patterns. (see fig. 12.8.) Which is the correct final position of the disks?

Choice "D" is the correct answer.

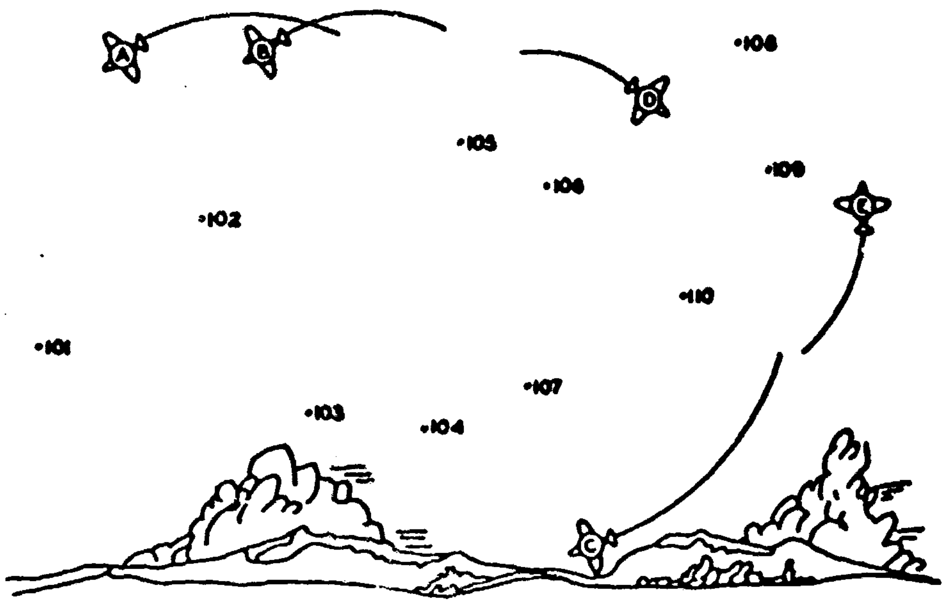


FIGURE 12.9
SAMPLE ITEMS OF FLIGHT PATH,
CP105A

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—None are available.

Evaluation of the Subarea of Visual Manipulation

The evidence presented in numerous factor studies for the existence of an independent visualization factor is substantial. All of the tests heavily saturated with the factor seem to involve a visual manipulative ability. In solving the problems, it is necessary mentally to move, turn, twist, or rotate an object or objects and to recognize a new appearance or position after the prescribed manipulation is performed.

Although it has been particularly difficult to construct pure measures of the ability, due to the ever present contamination with reasoning, there is little doubt that visualizing of the type described is involved.

Some progress has been made toward the understanding of this factor, but there is still much to be learned. No visualization factor, as such, had been extracted by previous investigators, although the possibility of its existence has been recognized for many years. The new tests, Position Visualization I and II and Spatial Visualization III, give promise of further defining and clarifying the concept.

The best estimate of the validity of the factor for pilots is 0.20, based upon all available results (see table 28.17). Estimates of validities for other air-crew specialties are 0.06 for navigators and 0.20 for bombardiers. Any test having a loading in the factor as high as 0.70 would thereby contribute 0.140 to pilot validity, 0.042 to navigator validity, and 0.140 to bombardier validity by reason of this factor alone.

VISUAL COMPLETION TESTS

Rationale of Visual Completion Tests

The typical test in this group calls for an ability to visualize the completion of a design or the extrapolation of a line or a path. This is merely another occasion for the visualizing abilities such as a pilot, particularly, seems required to bring to bear upon his job, as for example, he forecasts his own position and the positions of other airplanes, friend or enemy, perhaps in the next split second.

Flight Path, CP105A [19]

A pilot must be able to determine accurately his projected flight path. He must be able to judge beforehand relative positions of his plane and reference points along the planned course of flight. This test was an outgrowth of an earlier attempt known as the Line of Flight Test, CP102A,[20] which was abandoned because a single, uncontested extrapolation of the suggested irregular curves could not be determined.

Description.—In each item of the test the examinee must extrapolate an arc as he visualizes a plane in flight completing a circle. Only a part of the circular course of each plane is shown. Seeing only a part of the circle, the examinee must decide through which of several points the plane would pass if it continues along the same curve.

(1) *Internal characteristics.*—The directions include a page of five recorded, but unscored, sample items. Parts I and II each contain 3 pages of 10 items each, making a total of 30 scored items per part. Five

[19] Developed at Psychological Research Unit No. 3. Chief contributors: S/Sgt. J. Gordon Elkin and S/Sgt. Benjamin Fruchter.
[20] Suggested by Maj. Lewis B. Ward.

planes and their partial paths are illustrated on each page, with 10 lettered points through which the planes might pass.

(2) *Administration.*—Administration of directions and sample items consumes approximately 10 minutes, and 10 minutes are allowed for each of the 2 parts, making a total testing time of 30 minutes.

The sample items are shown in figure 12.9. Following are parts of the directions:

Each lettered plane is traveling a different circular path; that is it will make a complete circle. Your task is to decide which plane will pass through each numbered point. The points, which correspond to item numbers on your answer sheet, are numbered on the page from left to right (in the sample from 101 to 110).

If plane A followed the arc, it would go through point 101. Therefore, the answer to item 101 is A.

Each point has one and only one plane which will pass through it. Each plane will pass through one or more points.

Statistical results.—This test was administered at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 12.12. The distribution curves are approximately symmetrical and considerably flatter than normal.

TABLE 12.12.—*Distribution constants for Flight Path, CP105A*

Group	Scoring formula	Part	N	M	SD
Classified pilots[1]	Rights	I and II	1,332	29.9	10.0
Classified pilots[1]	Wrongs	I and II	1,332	24.0	9.8
Unclassified aviation students[2]	Rights	I	500	14.6	5.4
Unclassified aviation students[2]	Rights	II	500	15.4	6.8
Unclassified aviation students[2]	Rights	I and II	500	30.0	10.3
Unclassified aviation students[2]	Wrongs	I	500	11.3	4.9
Unclassified aviation students[2]	Wrongs	II	500	12.4	6.6
Unclassified aviation students[2]	Wrongs	I and II	500	23.7	10.5

[1] In class 44L.
[2] Tested in May 1944.

(2) *Reliability coefficient.*—For right and wrong scores separately, one sample yielded the estimates of reliability given in table 12.13.

TABLE 12.13.—*Estimated reliability coefficients for Flight Path, CI105A, based upon a sample of 500 unclassified aviation students[1] alternate-forms procedure*

Scoring formula	r'_{tt}	r_{tt}
Rights	0.46	0.63
Wrongs	.50	.67

[1] Tested in May 1944.

(3) *Correlation between rights and wrongs.*—The data are presented in table 12.14.

TABLE 12.14.—*Correlation between rights and wrongs for Flight Path CP105A*

Group	N	r_{rw}
Unclassified aviation students[1]	500	−0.50
Classified pilots[2]	521	−.44

[1] Tested May 1944.
[2] Tested May 9 to July 10, 1944.

703320—47—20

(4) *Test validity.*—Validation results based on two samples are given in table 12.15.

TABLE 12.15.—*Validity data for Flight Path, CP105A, based upon graduation-elimination of pilots in primary training*

Scoring formula	N_t	p_e	M_p	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
Rights	[2]1,332	0.86	30.07	28.64	10.04	0.08	0.14
Wrongs	[2]1,332	.86	23.94	24.53	9.78	−.03	−.09
Rights	[3]523	.75	32.29	30.31	10.48	.12	.16
Wrongs	[3]523	.75	21.46	24.58	9.38	−.20	−.23
R−W/4	[3]523	.75	26.96	23.62	12.08	.14	.21

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44I.
[3] Tested May 9 to July 10, 1944.

(5) *Item validity.*—Validation of items of this test disclosed the results recorded in table 12.16.

TABLE 12.16.—*Validity of items of Flight Path, CP105A, based upon samples of pilots in primary training*

N_t	p_e	$M\phi$	$SD\phi$	Range of ϕ	
				Low	High
[1]734	0.82	0.025	0.05	−0.12	0.14
[2]799	.86	.004	.04	−.07	.11

[1] In class 44H.
[2] In class 44I.

Evaluation.—Flight Path failed to exhibit promising validity for pilots. Chief interest in the test lies in the question of whether or not it will measure some new factor or factors not hitherto defined. A low communality with existent tests indicates promise in this direction. Before its factorial content, which appears to be largely specific in combination with presently available tests, can be revealed, other tests with similar content will have to be introduced into the correlational matrix to be analyzed. Low validity for pilots is not absolute proof that it is not measuring to a small extent the same visualization factor that manipulation tests have in common. Should it prove to be deficient in this factor, however, we have evidence for the hypothesis that the visualization factor is of a special variety, perhaps confined to manipulation tests.

Of all tests mentioned in this chapter, this one comes nearest from the test-constructor's viewpoint to satisfying the oft-mentioned ability "visualization of the flight course." From that aspect it has good face validity.

Evaluation for the Subarea of Visual Completion

Since Flight Path is the only test studied in this area, there is little evidence from which to draw conclusions. Flight Path itself has no correlation of more than 0.23 with any test on which sufficient data are available for analysis. Such low correlations suggest that its common variance would be so low that it would fail to appear significantly projected on any known factor. For that reason and because no apparently similar tests are known, it has not been used for analysis in factor studies. On an a priori basis, it would be easier to rationalize the presence of a perceptual factor, a distance-estimation factor, or a resistance-to-illusion factor than to explain why the known visualization factor should show saturation in the test. Obviously, flight path contains a substantial variance that is unique, as far as known tests are concerned.

EVALUATION OF VISUALIZATION TESTS

The questions quoted in the introduction to this chapter can be answered more satisfactorily now than before the present work was begun. In answering these questions, new problems have arisen which, it is hoped, will promote further research in the area.

The first question, "Is visualizing 'flat' forms the same as visualizing solid forms as they would appear from different sides?" is apparently answered satisfactorily. Thurstone admits of entertaining the hypothesis that two and three-dimensional spatial thinking might appear as two separate abilities, until the emergence of a single spatial-visual axis in his subsequent analysis denied the probability. Problems requiring the examinee to rotate flags, figures, cards, and lozenges in a flat plane appeared with factor patterns similar to those of problems involving three-dimensional manipulations, such as switching the surfaces of the objects visualized. Further evidence is presented in the data of this chapter to indicate that the important feature of visualization is not whether one, two, or three-dimensional movement of the visual image is concerned, but whether movement of any sort takes place. In the concept of visual manipulation a movement of some kind seems essential.

The second question referred to in the beginning of this chapter, "Is visualizing solid objects the same as visualizing movement of parts in a diagram of a machine?" can be answered with a positive "yes." In the aviation-psychology program, the best known measures of visualization are Spatial Visualization I (paper folding), Mechanical Principles, and Spatial Visualization II.

Question number three was, "Considering only flat forms, is the same ability required in visualizing several shapes singly as in visualizing how these shapes could be fitted together?" This question is not answered. According to the expressed description of manipulatory visualization, some movement is required, and visualizing several shapes singly seemingly requires none. By this token, only the manipulations involved in imagining how the shapes could be fitted together could be truly visualization of this sort. Then what is the ability to visualize these shapes singly? Since no test has been analyzed that can claim to measure such an ability, only preliminary hypotheses can be offered.

If the forms or details visualized are familiar, then possibly the ability involved is one of pure recall, and the visual-memory factor already reported will account for the variance. If, on the other hand, these shapes must be constructed or created in the "mind's eye," an entirely different ability may be needed. It is also possible that visual construction or completion might involve manipulations, insofar as each part or detail is "moved" into the visual picture and relegated to its proper position. If so, the ability could be explained by the recognized visualization factor. These are interesting questions and ones to which answers could be utilized to great advantage.

Several studies designed to seek objective answers to these questions were begun or were planned in the later stages of the war-time research program, but time did not permit their execution.

BIBLIOGRAPHY

(1) Thurstone, L. L. *Primary Mental Abilities*, Psychometric Monograph No. 1, University of Chicago Press, 1938.

(2) Kelley, T. L. *Crossroads in the Mind of Man*, Stanford University Press, 1928.

CHAPTER THIRTEEN

Mechanical Tests[1]

HISTORICAL STATEMENT

It has long been recognized by psychologists, vocational-guidance authorities, and others concerned with the most efficient employment of individual capabilities that differences exist in the abilities of individuals to succeed in pursuits involving the operation and utilization of mechanical equipment. Many attempts have been made to measure the ability or abilities involved and thus arrive at a reliable and valid basis for predicting success in such tasks. The resulting instruments may be classified roughly in three categories: (1) job-sample tests, (2) manual-ability tests, and (3) paper-and-pencil or printed tests. This chapter is concerned with printed tests. No extensive recapitulation will be made of results of previous civilian research in this area, but brief consideration of some of the outstanding studies may constitute a useful frame of reference within which to consider the results of the research reported in this chapter.

Some Standard Mechanical Tests

A number of printed tests of mechanical abilities have been available to the public for some time. Some of these will be mentioned briefly.

The Stenquist "Mechanical Aptitude Tests" (4) constitute one of the earliest attempts (1921) to measure mechanical ability or aptitude by means of printed tests. The two tests include mechanical-comprehension and mechanical-information material. Results on these tests are reported to correlate highly with ratings of proficiency given to students by shop and science teachers. As a first effort in the field of selection of workers for mechanical tasks, these tests are historically important.

Another milestone in the development of methods of measuring mechanical aptitude or ability was provided by the Cox (2) Mechanical Aptitude Tests (1928). Tests D and E consisted of material of the mechanical-comprehension and mechanical-information types. These tests were constructed in England and have been used with considerable success in that country in selecting individuals for mechanical tasks. The author recognized the differential effects of training and experience, and attempted to make the tests as independent as possible of these aspects.

The publication in 1936 of Thurstone's factor analysis of 57 tests (5), among which was a "Mechanical Movements" test, presented results of a

[1] Written by T/Sgt. Paul C. Davis.

new approach to the problem of analyzing and evaluating objective tests. Some discussion of this approach is found elsewhere in this volume. Unfortunately, no clear definition of the factorial composition of Thurstone's "Mechanical Movements" test was achieved, probably due to the fact that not enough "mechanical" tests were included in the matrix.

Using the factor-analysis method, Harrell (3) examined a group of mechanical tests and found five factors. These were identified as (1) verbal, (2) manual dexterity, (3) youth, (4) spatial, and (5) perceptual. Harrell concluded that the last two were the only ones that uniquely identify mechanical tests. A factor analysis to be reported in this chapter does not find manual dexterity, since only printed tests were involved. It does find spatial and perceptual variance, in confirmation of Harrell's results, but two other factors far outweigh those two in most printed mechanical tests.

The list of traditional printed tests designed to measure mechanical aptitude or ability includes the O'Rourke Mechanical Aptitude Test and the Mechanical Comprehension Test (Form AA by Bennett, Form BB by Bennett and Fry (1)). The O'Rourke test includes a part devoted to pictorial-comprehension items and another part containing verbally presented mechanical-information items. The Bennett tests consist entirely of pictorially presented, practical, mechanical or physical problems. Scores on these tests are reported to be positively correlated with success in shop work, with success in vocational training courses, and with the degree of complexity of mechanical tasks in which examinees were employed. Information is not available, however, as to their correlation with success in specific mechanical tasks. A test, similar to Mechanical Comprehension Test BB, was constructed by Bennett for the United States Navy for air-crew-selection purposes. High correlation was reported between scores on this test and success in pilot training.

Two Lines of Research Indicated

This brief review of research suggested two important aspects to be explored: (1) the relationship of mechanical tests to success in specific air-crew tasks, and (2) the factorial content of printed mechanical tests. This chapter reports results of research that has explored these two areas to a significant extent.

MECHANICAL ABILITY AS RELATED TO AIR-CREW PERFORMANCE

Job Analysis Findings

It is generally recognized that successful performance of tasks involving the use or operation of mechanical devices requires certain special abilities. It has been assumed by some who have been interested in the problem, however, that ability to succeed in such pursuits depends upon factors not unique to mechanical tasks. On the basis of such an assumption, job analyses of tasks involving use of machines might not even include reference to machines, or their operation, as such.

In general, this seems to be true of the various job analyses of air-crew duties, since the reports typically fail to mention mechanical abilities, at least under this rubric. No explanation is made for this omission, but the reasoning of the preceding paragraph probably identifies one answer. Another explanation might be that the levels of mechanical ability required are low enough to accommodate most or all who reach the training stage.

Mechanical Requirements for Air Crew

The opposite hypothesis, that mechanical ability as such is unique and is an important determiner of success in air-crew training, has been adopted by some. A brief review of the mechanical content of the jobs of pilot and bombardier will reveal, in part, the basis of this hypothesis.

The pilot.—The pilot of a military plane is in full charge of its operation with responsibility for its proper functioning and, in the case of a bombing plane, for the safety of the entire crew. In view of these responsibilities, the pilot must perform certain duties on the ground prior to take-off. These duties include thorough inspection of the plane to determine whether it is in proper condition for the take-off and for safe operation in the air. At first in ground school and later in other phases of his training, the pilot is instructed in the construction and function of the airplane and is trained in the meticulous performance of all duties related to its operation. The pilot obviously must understand well the mechanical parts and functions of the airplane in order to make an intelligent and exacting inspection.

While the plane is in flight, the pilot is faced with the task of observing, interpreting, and acting upon information received from dials, indicators, etc., in the plane. Especially under combat conditions, the pilot may frequently be forced to supervise improvisation of repairs on damaged equipment or to devise means of replacing destroyed parts of the plane. Such emergency action requires a clear grasp of practical mechanics. The assumption that the ability to devise or improvise under these conditions varies greatly is probably justifiable.

The bombardier.—The mechanical aspects of the bombardier's task are principally related to the operation of the bomb sight. Early in his training, the bombardier begins a study of the bomb sight. By the end of his training, the bombardier has studied every part of the bomb sight and is equipped to diagnose malfunctions, and to make minor repairs. Before every training or combat flight, the bombardier "preflights" the sight. This operation includes checking the functioning of all parts of the sight, calibrating indicators, and setting the sight for constant data, such as field elevation and the like. In addition to the care of the bomb sight, the bombardier has full responsibility for loading and arming of the bombs and inspecting of bomb racks in the plane. Thorough knowledge of their proper functioning is necessary to enable the bombardier to insure proper and safe operation of racks and bomb release mechanism on the mission.

Principal flight duties of the bombardier include operation of the bomb sight and the automatic-pilot equipment. Ordinarily these tasks consume only a few minutes, but the entire success of the mission depends largely upon the accurate performance of these tasks. In emergency situations, the bombardier may be called upon to improvise a method of releasing the bombs, opening bomb bays, or the like.

Measurement of Mechanical Ability

From these descriptions of pilot and bombardier duties, it may be seen that the tasks involved include a great deal of mechanical content. Establishment of this fact is, however, only the first step in the process of determining aptitude for the tasks. Of equal importance is the manner in which the specific ability is to be measured.

Why printed tests were utilized.—The problem of measuring nonintellectual or partially intellectual abilities by means of printed tests has ever been a difficult one. The area of mechanical ability is no exception, and many are prepared to claim that adequate measurement of this ability is not possible by such techniques. In the early days of psycological testing in the Army Air Forces, however, it was imperative that means be sought to measure as many aptitudes or abilities as possible by paper and pencil methods. This was due largely to the fact that adequate psychomotor or job-sample tests did not exist in many areas and the pressure of great numbers of examinees made the use of printed group tests wherever possible highly desirable.

There are adequate reasons why this method of testing should prove successful. On close examination, many or most practical mechanical problems prove to have an important intellectual component as distinguished from purely manipulative skill. This component is probably not strictly "abstract" intelligence and certainly is not the same as verbal ability. It includes the ability to gain insight into the principles involved in mechanical problems. A second reason is that it logically may be assumed that mechanical insight will result in or tend to result in solution of practical mechanical problems.

Causes of Individual Differences in Mechanical Abilities

Owing to the heterogeneity of the group of prospective air-crew members, it is necessary to consider the factor of differential mechanical experience. Apparently, unusual amounts of information in the field of mechanics may stem from one or more of three causes: (1) the tendency or desire to seek mechanical experience, (2) superior ability to profit by mechanical experience, and (3) unusually rich opportunity to gain mechanical experience.

Scores on mechanical tests may reflect individual differences in all three of these aspects, depending upon emphasis. Mechanical interest may be prognostic in that the air-crew job undertaken promises satisfaction of such an interest. Ability to profit by mechanical experience is also a favorable trait for learning the mechanical aspects of flying. Generous opportunity for mechanical experience would be of value only insofar as it resulted in habits or information that transfer to air-crew performance. A student who makes a high mechanical-test score because of unusual opportunity is probably a less good risk for training than one whose opportunity may have been limited but whose aptitude for mastering mechanical tasks is high.

The first two features—interest and aptitude—are probably positively related. To measure the one is thus to measure the other to some extent. Since they are both probably favorable to success, their intermixture is not a serious matter.

Unusual opportunity to gain mechanical experience would probably not correlate appreciably with either superior aptitude for things mechanical or with the tendency to seek mechanical experience. One would therefore attempt to minimize the variance in opportunity in favor of variances in one or the other of the first two. As it turns out, measurements represent an intermingling of the three, and one can only trust that the prognostic components are not too much submerged for practical purposes.

The Plan of Research

In view of the apparently complicated nature of mechanical ability, it appeared advisable to employ several approaches. Preliminary "armchair analysis" indicated two types of measures which should be useful: measures of mechanical comprehension and measures of mechanical information. The original plan included also measures of pattern comprehension because such measures had been traditionally included in the mechanical area. These tests were found to have little in common with the mechanical tests, however, and are described elsewhere in this volume. A test of physics was constructed for another purpose, but because of its close relationship (superficial, at least) to mechanics, it is treated in this chapter.

MECHANICAL COMPREHENSION

Definition and Rationale

The area covered by mechanical comprehension is broad and includes a wide variety of possible approaches. In general, mechanical comprehension may be defined as the ability to follow, to understand, and to predict the result of the operation of mechanical devices or machines. Results obtained from the observation of actual machines would probably yield the most valid measures, but two-dimensional pictorial representations appear to be a fair substitute for actual machines. This substitution is based upon the assumption that solution of two-dimensional problems, where inspection but not manipulation is possible, involves the same or similar intellectual functions as those required in solving problems with three-dimensional machines. Variety can be secured by presenting machines at various levels of complication, ranging from the simple one- or two-part machine to the engine that contains a multitude of parts.

Another approach explores knowledge or comprehension of common physical laws encountered in everyday, nontechnical experience. Good rationale for utilizing this approach is found in the fact that the problems used may include material common to the experience of all or almost all examinees. This fact should tend to minimize the effects of differential experience. Additional justification of this approach lies in the fact that principles involved in complicated machines are also involved in much simpler and more fundamental form in mechanical or physical phenomena of everyday life.

Mechanical Principles, CI903A [3]

An understanding, at least in a naive manner, of basic principles governing mechanics appears to be fundamental to the solution of even relatively simple specific mechanical problems. Obviously, it would be impracticable to construct a large number of tests, each designed to explore but one principle involved in such an activity as flying, even if it were possible to isolate the principles and to prove their pertinence. A single test which would explore the examinee's familiarity with a large number of these basic principles, however, appeared to be feasible. Such a test was constructed and given the appropriate title, "Mechanical Principles Test."

A preliminary form of this test (CI903AX) constituted the exploratory instrument upon which Form CI903A was based. A study of Bennett and Fry's "Mechanical Aptitudes Test," Thurstone's "Mechanical Movements Test," and other similar tests was made in preparing the preliminary form. An effort was made to give the test face validity by introducing practical mechanical principles in terms of aviation situations whenever possible. Eighty-six items were drawn up and administered experimentally. On the basis of item analysis, the 30 items yielding the highest internal-consistency phis, and at the same time the most appropriate difficulty indices, were selected to be used in Form CI903A.

Description.—Mechanical Principles Test (CI903A) consists of 2 sample practice items and 30 scored items, all presented pictorially. In each item, the examinee is asked to select the answer that describes most accurately what is happening or will happen in the pictured situation. The sample items in figures 13.1 and 13.2 are typical. The problem in

[3] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter Jr., Capt. S. W. Cook, T/Sgt. Paul C. Davis, Lt. Lina Hutchinson.

figure 13.1 is to determine which plane is about to turn to the left. In figure 13.2, the examinee is required to determine which hook, if either, is capable of lifting the heavier weight.

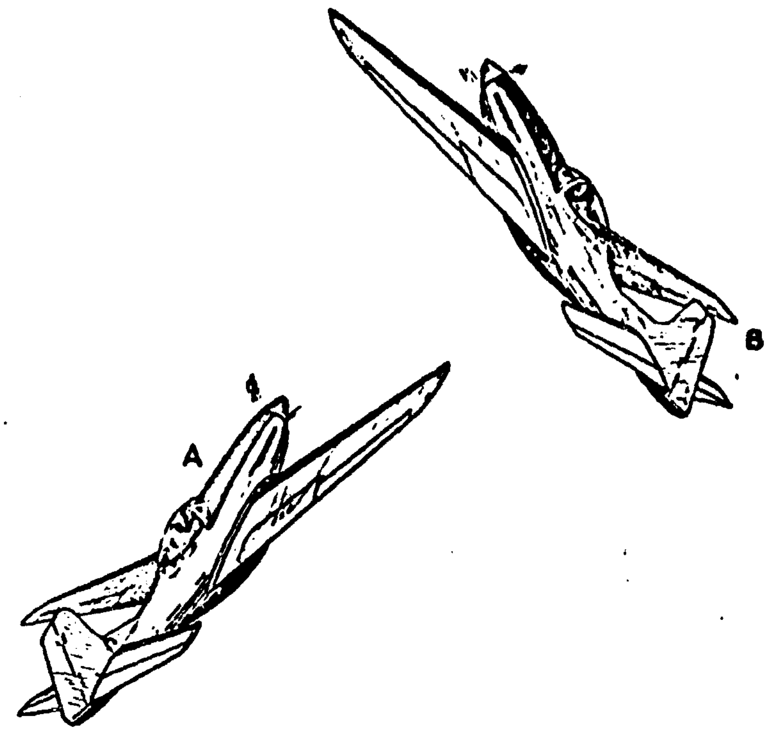


FIGURE 13.1
SAMPLE PROBLEM OF MECHANICAL PRINCIPLES, CI903A

2. Which plane is about to turn left?

   2–A Plane A
   2–B Plane B
   2–C Both plane A and B

(1) *Internal characteristics.*—The examinee is required to choose one of three alternatives. In most cases, the third alternative is a midpoint or neutral position between the other two alternatives, such as "equal," "the same," "either," "neither," "both," etc. Unfortunately, it was impossible to construct the test in such a way as to produce as many correct answers in the neutral categories as in each of the other two. Twenty-eight items contain third alternatives of this type, of which only five are keyed as correct.

(2) *Administration.*—Instructions to the examinee are very simple, being presented principally in connection with two sample items. Three or 4 minutes suffice for the directions, and 15 minutes are allowed for the scored items. Although rapid work is required to complete the test in this time, 80 percent to 90 percent of the examinees are able to finish.

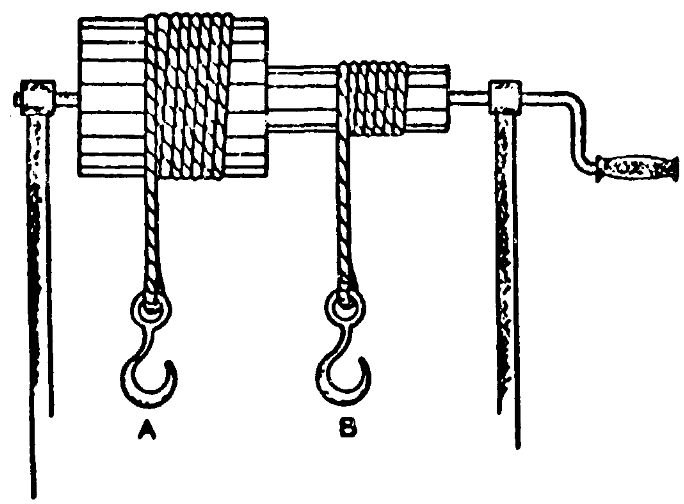


FIGURE 13.2
SAMPLE PROBLEM OF MECHANICAL PRINCIPLES, CI903A

30. On which hook could you lift the heavier weight?

30–A Hook A
30–B Hook B
30–C Equally heavy on both

(3) *Scoring.*—In view of the fact that very few examinees choose the third alternative, formulas, first, of R–W and, later, of 2R–2W were used in scoring the test. These formulas, however, produce a large number of negative scores, which are rather difficult to handle in combining them by machine with other scores, so formulas of 2R-2W+20, 2R-2W+30, and 2R-2W+40 were successively employed in order to secure better adaptation to aggregate-weighting requirements.

Statistical results.—Numerous statistics are available on this form of the test.

(1) *Distribution statistics.*—Table 13.1 presents distribution data for several samples.

TABLE 13.1.—*Distribution statistics for Mechanical Principles, CI903A, based upon samples of unclassified aviation students*

Psychological research unit No.	Testing date	Score	N	M	SD
3	August 1942	R−W	359	8.1	9.6
1	December 1942	2R−2W	1,096	18.6	15.9
2	December 1942	2R−2W	1,015	18.4	15.8
3	December 1942	2R−2W	1,143	18.8	16.0
1, 2, 3	July 1943	2R−2W+40	3,000	58.6	18.7

(2) *Optimal scoring formula.*—A study to determine the optimal scoring formula to maximize pilot validity showed that the rights and wrongs should be weighted 1.00 and +0.18, respectively. This yielded an estimated validity of 0.47 for primary pilot training. The data from

which this weighting was derived are based on a sample of 1,094 students in pilot training in class 43H, originally tested at Psychological Research Unit No. 2. The means for rights and wrongs scores were 18.8 and 10.3, respectively. The intercorrelations were: $r_{RC}=0.47$, $r_{WC}=-0.41$, and $r_{RW}=-0.91$, in which subscript "C" refers to the criterion. For practical scoring purposes, a weight of zero is recommended for the wrongs score when the test is used for the selection of pilots. In view of the almost perfect negative (-0.91) correlation between rights and wrongs, however, a weight of -1 which was used in scoring would produce validity only slightly lower than that estimated for the recommended weight of zero.

(3) *Internal consistency.*—An item analysis, based upon the performance of the highest 95 (25 percent) and lowest 95 of 380 unclassified aviation students tested at Psychological Research Unit No. 3 yielded internal-consistency phi coefficients ranging from 0.22 to 0.67. The mean phi value was 0.44, with a standard deviation of 0.12. As indicated by the data, the items proved to be quite homogeneous. This consistency was achieved in large part by selecting from the 86-item preliminary form the items with the highest phi's and the most appropriate difficulty.

(4) *Reliability coefficients.*—Several reliability estimates are given in table 13.2.

TABLE 13.2.—*Reliability coefficients for Mechanical Principles, CI903A*

N	Type	r_{tt}
[1]500	Odd-even	0.76
[2]240	Odd-even	.75
[2]240	Odd-even	.69
[2]255	Kuder-Richardson IV	.73
[2]255	Test-retest[3]	.90

[1] Tested January 23, 1943, at Psychological Research Unit No. 2.
[2] Tested April 1943 at Psychological Research Unit No. 3. Non-overlapping samples.
[3] Test-retest interval not reported.

(5) *Difficulty.*—Based upon the analyses referred to previously, the mean difficulty of the items, corrected for chance, is approximately 0.55, which is considered to be satisfactory.

(6) *Factorial composition.*—The Mechanical Principles test was included in several factor analyses. Two of these contained principally mechanical tests plus a few tests from the classification battery. These analyses agree in the main in accounting for the variance of the test. The highest loading occurs in a mechanical-experience factor, its mean

TABLE 13.3.—*Preliminary validity of Mechanical Principles, CI903A, for success in primary pilot training*

N	p_s	M_s	M_u	SD_t	r_{bis}
[1]302	0.77	11.10	5.72	9.12	0.35
[2]314	.88	13.04	7.64	8.62	.33

[1] In classes 43E and 43F. Tested August 17, 20, and 21 at Psychological Research Unit No. 3.
[2] Tested November 25 and 26, 1942 at Psychological Research Unit No. 3.

Table 13.4.—*Pilot validity of Mechanical Principles, CI903A, using the graduation-elimination criterion*

Group	N	Class	Psychological research unit No.	Score	p_s	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training	652	43H	2	2R−2W+40	0.59	62.6	46.4	19.8	0.51	....
Pilots in primary training	910	43H	3	R−W	.81	13.0	7.0	9.4	.36	....
Pilots in primary training	2,100	43J	1, 2	2R−2W+40	.63	61.2	51.8	17.9	.32	....
Pilots through advanced training	1,990	43J	1, 2	2R−2W+40	.52	61.9	53.0	17.9	.31	0.37
Pilots in primary training	4,779	44E	1, 2, 3	2R−2W+40	.88	62.3	50.5	18.2	.34	.37
Pilots in B-17 transition training	1,045	43J, K	1, 2, 3	2R−2W+40	.98	63.3	58.9	17.7	.10	....
Pilots in B-24 transition training	980	43J, K	1, 2, 3	2R−2W+40	.92	63.4	56.7	17.1	.19	....
Pilots in B-25 transition training	313	43J, K	1, 2, 3	2R−2W+40	.98	63.4	70.6	16.5	−.19	....
Pilots in B-26 transition training	376	43J, K	1, 2, 3	2R−2W+40	.82	63.2	60.7	17.9	.08	....
Pilots in P-40 assignment	2,416	43J, K	1, 2, 3	2R−2W+40	.64	62.4	61.7	17.0	.03	....
Pilots in primary training	1,951	43J	3	(2)	.74	5.5	4.6	2.0	.27	....
Pilots in primary training	2,495	43K	3	(2)	.81	5.6	4.9	1.9	.26	....

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Using scaled scores with a mean of 5.00 and a standard deviation of 2.00.

Table 13.5.—*Validity of Mechanical Principles, CI903A, using the criteria of graduation-elimination from and grades in navigation training*

Criterion	N	Class	Psychological research unit No.	Score	p_s	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$
Graduation-elimination	410	43-10, 11	1, 2	2R−2W	0.85	18.3	9.2	15.8	0.32	[1]0.39
Graduation-elimination	1,326	43-12 through 43-15	1, 2, 3	2R−2W+40	.80	54.4	51.1	19.0	.10	[3].20
Grades in dead reckoning (ground school)	224	43-10 through 43-15 (Hondo)	1, 2, 3	....	....	....	....	....	[2].17	[3].27
Grades in celestial navigation (ground school)	....	....	....	....	....	....	....	....	[2].07	[3].17
Grades in dead reckoning (flight)	....	....	....	....	....	....	....	....	[2].05	[3].12
Grades in celestial navigation (flight)	....	....	....	....	....	....	....	....	[2].14	[3].22
Grades in meteorology	....	....	....	....	....	....	....	....	[2].21	[3].29
Military grades	....	....	....	....	....	....	....	....	[2].05	[3].09
Final composite grades	....	....	....	....	....	....	....	....	[2].14	[3].25

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Product-moment correlation.
[3] Assumed unrestricted stanine standard deviation not reported.

TABLE 13.6.—*Validities of Mechanical Principles, CI903A, for bombardiers in training*

Group	Criterion	N	Class	Psychological research unit No.	Score	p_c	M_c	M_e	SD_t	r_{bis}	$_c r_{bis}$
Bombardiers[1]	Graduation-elimination	953	43-8 through 43-11	1, 2, 3	2R−2W	0.79	11.2	9.0	13.1	0.10	0.12
Bombardiers[2]	Graduation-elimination	439	43-14 through 43-18	1, 2, 3	2R−2W+40	.85	46.0	42.6	16.7	.11	.14
Bombardiers[3]	Graduation-elimination	216	43-14 through 43-18	1, 2, 3	2R−2W+40	.85	56.7	50.6	19.0	.17	.19

[1] New aviation cadets taking 12-week course (no navigation training).
[2] New aviation cadets taking 18-week course (some navigation training).
[3] Reclassified pilots taking 18-week course.

loading in the various analyses being 0.60. The second highest loading (0.51) is in a visualization factor. Other significant loadings are in the spatial-relations (0.22), verbal (0.20), and general-reasoning (0.20) factors. Its communality is found to be 0.84 when all factorial results are summarized.

(7) *Test validity.*—Preliminary validation data, gathered on two different samples of aviation students in which little selection had been made by disqualification at the time of classification, are shown in table 13.3. As indicated by table 13.3, there was good evidence of the validity of the Mechanical Principles test for pilot training. The test became a part of the classification battery of 1 December 1942, and validation data were subsequently accumulated for all three air-crew positions and for certain other specialties. Data for several samples are given in tables 13.4 through 13.7.

TABLE 13.7.—*Validities of Mechanical Principles, CI903A, for combat crew specialties*

Group	N	Criterion	r[1]
Air mechanics[3]	213	Average grades	0.42
Armorers[2]	258	Average grades	.01
Flexible gunners[4]	194	Air-to-air firing[5]	−.02
Flexible gunners[4]	194	Final examination	.35
Flexible gunners[6]	173	Air-to-air firing[5]	.11
Flexible gunners[6]	173	Final examination	.46
Radio operator mechanics	235	Graduation-elimination	−.02
Radio operator mechanics	153	Average grades	.27

[1] Product moment correlation.
[2] Tested with the December 1942 Battery at Psychological Research Unit No. 2.
[3] Tested with the December 1942 Battery at Psychological Research Units Nos. 1, 2, and 3.
[4] In class 43-45 at Ft. Myers. Tested at Psychological Research Units Nos. 1, 2, and 3.
[5] A very unreliable criterion.
[6] In class 43-48 at Ft. Myers. Tested at Psychological Research Units Nos. 1, 2, and 3.

As indicated by the validation data, the Mechanical Principles test shows greatest promise as a pilot-selection instrument. The biserial correlations against the graduation-elimination criterion in primary training of the samples given, combined by means of Fisher's z, is 0.33. The correlation with graduation-elimination from bombardier training (combined by the same method) is slight (0.13). The correlation with navigator training success (combined by the same method) is moderate (0.25).

(8) *Item validity.*—Subsequent to the inclusion of the Mechanical Principles test in the classification battery, pilot item-validity studies were made. The results are reported in table 13.8. The phis of the two samples correlate 0.68. A study of the relationship between internal-consistency phis and item validities revealed a marked positive correlation (0.64). This relationship strongly supports the practice of selecting items for the test on the basis of internal-consistency phi values.

When early data indicated that the validity of Mechanical Principles was considerable, while that of Physics was very slight for pilots, it appeared desirable to examine the items of the Mechanical Principles test

TABLE 13.8.—*Validity of items of Mechanical Principles, CI903A, for primary pilot training, graduation-elimination criterion*

N_t	p_t	$M\phi$	$SD\phi$	Range of ϕ	
				Low	High
1,094[1]	0.84	0.13	0.05	0.07	0.22
960[2]	.81	.11	.06	−.03	.24

[1] Tested Sept. 1 to Nov. 30, 1942, at Psychological Research Unit No. 2.
[2] Tested in Oct. 1943 at Psychological Research Unit No. 3.

with the view to eliminating those items that are highly correlated with Physics and are at the same time low or moderate in pilot validity. Since it was supposed that the total score on mechanical principles reflected some variance in knowledge of physics, it was recognized that the total score on Mechanical Princilpes was not the best criterion for internal-consistency studies. It was proposed that the total score on Mechanical Information would be an excellent criterion, because it was presumably free from physics variance. The Mechanical Principles items were therefor analyzed, using scores in both the Mechanical Information and the Physical Principles tests as criteria. Examination of the two sets of phi values from these analyses revealed so strong a positive relationship between them, however, that any selection on the proposed basis was difficult to justify.

At a later time, when factorial composition of the Mechanical Principles test was better known, it was desired to segregate items into separate pools, each relatively pure with respect to one of the factors. It was assumed that some items were strongly mechanical, others spatial, and still others visualizing items. The Mechanical Information, the Complex Coordination, and the Pattern Comprehension tests were chosen as criteria for the mechanical, spatial, and visualization factors respectively. The three sets of phi coefficients were again highly intercorrelated, making it impossible to sort the items into three factor categories, as had been intended.

Although it is recognized that the criterion tests were not pure measures of the factors they represented, it is apparent that items of the Mechanical Principles test do not fall into factor categories but that they are typically complex factorially.

Evaluation.—As indicated by the data presented, the Mechanical Principles Test, CI903A, proved to be one of the most useful pilot-selection instruments available. It combines in its total variance large amounts of two highly valid factors—mechanical experience and visualization—plus a smaller amount of another highly valid factor—spatial relations. Its nonerror variance and its validity are fully accounted for by known common factors.

Although proof of validity for air-crew success alone is available, the factorial findings strongly suggest the possibility of using such a test in selecting for a wide variety of mechanical pursuits. Its only defect is its

703320—47—21

lack of purity. When a strong test of visualization is available, this test could well be replaced with a combination of the visualization test and Mechanical Information. Each component could then be appropriately weighted, depending upon the criterion one desired to predict.

Mechanical Principles, CI903B [3]

This form of the test is the successor to Form CI903A already described. It appeared that revision and introduction of new material might provide a test which would be even more valid. Approximately 200 items, potentially useful in a new form, were designed and evaluated carefully against the following requirements:

(1) Apparent relationship to the more valid items in Form A.

(2) Minimum involvement with physics information and reading comprehension.

(3) Moderate difficulty (around 0.50).

(4) Adequate number of alternatives (preferably five).

From these 200 items, 120 were selected on the basis of the above requirements. These 120 items were separated into two groups of 60 each, as nearly comparable in difficulty and content as possible. These two groups became Forms CI903BX1 and CI903BX2 and contained identical instructions and illustrative items. The two forms were administered to 1,920 classified pilots and the results employed in making an item analysis. The upper and lower groups were determined by total scores on the two forms combined, since both forms were given to the same examinees. On this basis, internal-consistency phis varied from 0.00 to +0.51 on BX1, and from −0.02 to +0.55 on BX2. Twenty-four items in BX1 and twenty-three in BX2 yielded phis of +0.35 or above. In view of the high correlation (0.64) previously found between internal-consistency phis and item validities on Form A, major weight was given to internal consistency in selecting items from the BX1 and BX2 forms. Other considerations were difficulty and the functioning of misleads. Because of the high validity of several items in Form A, 14 items were selected to be redrawn and used in Form B. Twelve items from BX1 and 14 items from BX2 completed the group selected for Form B. Unfortunately, time did not permit the obtaining of validity data on the items in BX1 and BX2 before this selection of items was made.

Description.—This form of the Mechanical Principles test contains 40 scored items of the same general type as those in Form A. The instructions are similar to those for Form A.

(1) *Internal characteristics.*—The number of misleads ranges from three to five and averages four per item. There was a strong attempt to make all misleads functional, with the result that the alternatives in this form are much more uniform in appeal than are those in Form A.

[3] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, S/Sgt. Benjamin Fruchter, S/Sgt. Wayne S. Zimmerman.

(2) *Administration.*—The one sample item and directions require about 5 minutes. The test time allowance is 20 minutes.

(3) *Scoring.*—The test was first scored R−W/2, but this was later changed to R−W/2+20 to eliminate negative scores.

Statistical results. (1) *Distribution statistics.*—Scored with the formula R−W/2+20, the test yielded the data given in table 13.9.

TABLE 13.9.—*Distribution of scores on Mechanical Principles, CI903B*

Group	N	M	SD
Unclassified aviation students[1]	1,500	32.3	9.3
Unclassified aviation students[2]	1,920	30.4	9.0
Classified pilots[3]	3,146	35.4	8.6
West Point cadets[4]	888	35.5	9.9

[1] Tested November 1943 at Psychological Research Units Nos. 1, 2, and 3.
[2] Tested November 1943 at Medical and Psychological Examining Units No. 4 through 10.
[3] In class 44I. Tested November 1943 at Psychological Research Units Nos. 1, 2, and 3.
[4] Class of 1946.

(2) *Internal consistency.*—Internal-consistency phi values are available from Form A, BX1, or BX2, on all items in Form B. The phi values of the items in their original tests have a mean of 0.46, while the mean phi of the items as obtained by analysis of Form B is 0.48. The results of the item analyses are given in table 13.10. The two sets of phis correlate 0.47.

TABLE 13.10.—*Comparison of internal-consistency phi values obtained from Mechanical Principles, CI903B, and preliminary forms*[1]

Analysis	Mφ	SDφ	Range of φ	
			Low	High
Preliminary forms[2]	0.46	0.08	0.36	0.73
Form B[3]	.48	.11	.28	.70

[1] For samples of unclassified aviation students tested in October 1943 at Psychological Research Unit No. 3.
[2] Using upper 50 percent and lower 50 percent of 960 unclassified aviation students.
[3] Using upper 25 percent and lower 25 percent of 800 unclassified aviation students.

(3) *Reliability coefficients.*—Two estimates of the reliability of the test are given in table 13.11. The correlation between the two experimental forms (BX1 and BX2) is 0.70, based upon 970 cases. The best estimate of reliability may lie between the two limits of 0.70 and 0.83, but the communality (0.93) strongly suggests that the higher figure is nearer the correct value.

TABLE 13.11.—*Reliability coefficients for Mechanical Principles, CI903B, based upon samples of unclassified aviation students*

N	Type	r'_{tt}	r_{tt}
[1]500	Odd-even	0.69	0.82
[2]1,000	Alternate forms	.71	.83

[1] Tested at Medical and Psychological Examining Unit No. 10 with the November 1943 battery.
[2] Tested at Medical and Psychological Examining Unit No. 7 from Jan. 30, 1944, to Feb. 14, 1944.

(4) *Difficulty.*—In order to establish norms, Form B was administered to 530 unclassified students in October 1943 at Psychological Re-

search Unit No. 3. This sample yielded a mean score of 13.4 and a standard deviation of 9.1. The mean difficulty index, corrected for chance, was 0.30. In subsequent samples of classified pilots, however, this figure rose to 0.44, which is similar to the difficulty level of Form A.

(5) *Factorial composition.*—Although Form B of Mechanical Principles was in the classification battery for a longer period than Form A, it did not appear in so many factor analyses, and consequently, somewhat less is known concerning its factorial composition. Enough information is available, however, to indicate close factorial similarity to Form A. The mechanical-experience and visualization factor loadings remain approximately the same, 0.58 and 0.54 respectively, while the loading in the spatial-relations factor is significantly lower (0.12) than that for Form A (0.22). The loading in the verbal factor (0.03) is much lower than the 0.20 for Form A, and the loading in the general-reasoning factor (0.34) is markedly greater than in Form A (0.20). It was concluded, however, that little change of importance had taken place in the factorial content of the test in the change from Form A to Form B. The communality for Form B reached the unusual level of 0.93. The gain over that for Form A is accounted for by two new factors in which Form B has loadings—carefulness (0.17) and space III (0.28)—factors that did not appear in analyses including Form A.

(6) *Test validity.*—Preliminary statistics on the internal consistency of items selected for Form B indicated that higher validity could be expected for this form than was obtained with Form A (if the relationship between internal consistency and item validity found in Form A ($r = 0.64$) also prevailed in Form B). Unfortunately, from the standpoint of predicting validities in this manner, the total-score validity for Form B is about the same (0.34) if not slightly lower.

Subsequent to the inclusion of Form B in the classification battery, validation data were gathered for 3,146 pilots and two classes of WASP trainees. These data are given in table 13.12.

TABLE 13.12.—*Validity of Mechanical Principles, CI903B, for various aviation trainees*

Group	Criterion	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$
Pilots in primary training[1]	Graduation-elimination	3,146	0.84	35.95	32.39	8.56	0.23	[2]0.33
Pilots in primary training[3]	Graduation-elimination	1,676	.89	35.91	32.55	7.97	.22	[4].35
WASPs[5]	Graduation-elimination	91	.80	24.41	22.22	6.18	.20	....
WASPs[5]	Graduation-elimination	104	.61	23.60	20.00	6.70	.33	....
Field armament	Final grades	269	....	.....	.....	....	[6].04	....
Air mechanic	Final grades	254	....	.....	.....	....	[6].33	....

[1] In class 44I. Tested at Psychological Research Units Nos. 1, 2, and 3 with the November 1943 battery.
[2] Assuming an unrestricted stanine standard deviation of 1.90.
[3] In class 44J. Tested at Psychological Research Unit No. 3 with the November 1943 battery.
[4] Assuming an unrestricted stanine standard deviation of 1.83.
[5] WASP is the abbreviation for Women's Auxiliary Service Pilots. 91 cases in class 44-W-7; 104 cases in class 44-W-8.
[6] Product-moment correlation.

The corrected pilot validities of 0.33 and 0.35 were achieved notwithstanding a general decrease in validity of classification instruments, as typified by the Complex Coordination Test which suffered a reduction in validity from approximately 0.36 to approximately 0.32, between the July 1943 and November 1943 batteries. The validities found for the two small samples of WASP trainees indicate that the predictive value of the test is not limited to the male sex.

(7) *Item validity.*—A sample of 704 primary pilot trainees in class 44H, 600 of whom graduated, yielded a mean validity phi of 0.08, with a range of values from −0.01 to 0.16, and a standard deviation of 0.05.

Evaluation.—Like Form A of Mechanical Principles, Form B proved to be one of the most valid instruments used in selecting pilot trainees for the Army Air Forces. The efforts to produce a test more valid than Form A had apparently failed. With due allowance for a changed criterion, as mentioned before, a part of the failure was due to the reduced loading in spatial relations.

The predicted validity for Form B is 0.31 (see table 28.18). The loss of spatial variance was not a serious matter, since it was covered by other tests. Other available data, such as validity for air mechanic trainees, suggest wider usefulness for this test.

Item-validity data indicate that selection of items to yield maximum pilot validity had probably not yet been achieved. While the test is also valid in the selection of women pilots, its value for this purpose is somewhat less than in the case of men, due, no doubt, to the narrower range of ability (probably the mechanical-experience component) as indicated by smaller standard deviations (see table 13.12).

Mechanical Functions, CI907AX [4]

This test was constructed for the purpose of measuring (1) knowledge of tools and instruments and (2) ability to comprehend the method of operation of machines. The latter was conceived as a rather complicated function covering, among other things, an integration of specific mechanical abilities, such as those measured in the Mechanical Principles, Mechanical Information, and Mechanical Movements tests. The understanding of the operation of individual parts is probably necessary to an understanding of the operation of a machine. Such specific knowledge or ability, however, may not insure understanding of the operation of the machine as a whole.

Description.—This test is constructed in two parts that are quite dissimilar. Part I consists of items showing pictures of tools that are to be identified. This part was subsequently revised and became the Tool Function test. In some of the problems, a single tool is pictured, and the examinee is required to identify its use from a list of alternatives. Other problems picture five tools from which one is to be selected as having a

[4] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter, Jr., T/Sgt. Paul C. Davis, Lt. Linn Hutchinson.

certain specified characteristic or use. The problem shown in figure 13.3 is typical of the items in this part. In this problem, the examinee is re-

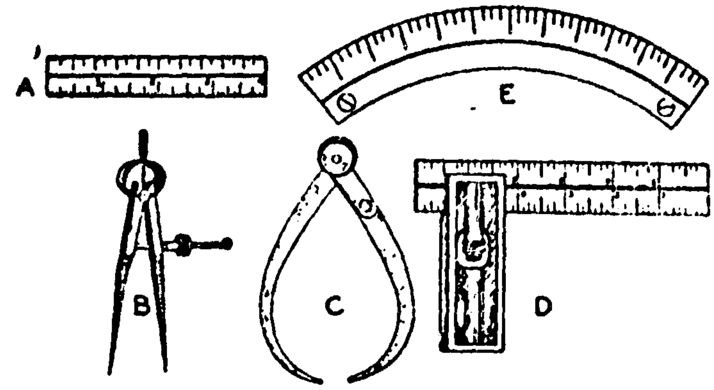
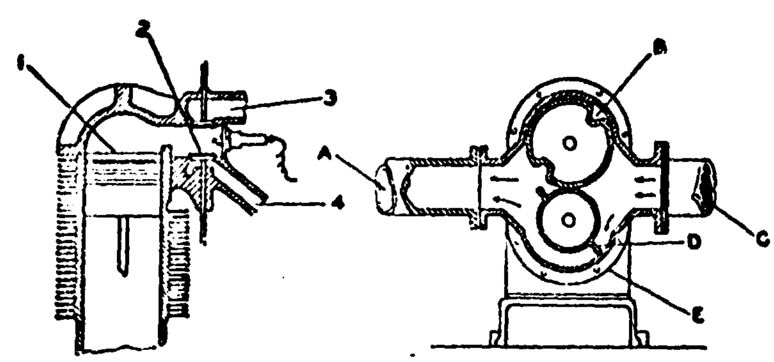


FIGURE 13.3
SAMPLE PROBLEM OF PART I, MECHANICAL FUNCTIONS, CI907AX

quired to identify the tool that is best for measuring outside work on a lathe.

Part II consists of items showing pictures of machines of varying degrees of complexity. The problems consist of identifying, from a list of alternatives, the functions of either the whole machine or of certain specified parts. Several of the problems require that this identification be made in terms of analogous parts of two different machines. A typical item is shown in figure 13.4. The examinee is required to discover what



AIR COOLED CYLINDER A DRUM PUMP

FIGURE 13.4
SAMPLE PROBLEM OF PART II, MECHANICAL FUNCTIONS, CI907AX

parts of the drum pump do the same things as parts 1 and 2 of the air-cooled cylinder.

(1) *Internal characteristics.*—There are 15 items in each part.

(2) *Administration.*—In view of the simplicity of the task in this test, directions are limited to general instructions concerning the method

of answering, guessing, and timing. These instructions require about 3 minutes. Six minutes working time is allowed for part I, and 10 minutes for part II.

(3) *Scoring.*—The test is scored with the formula R−W/4. Part scores were computed and used exclusively in view of the dissimilarity of the parts.

Statistical results.—This test is one of those included in a battery of mechanical tests, and rather complete statistics are available, although the samples are not large. The data are for unclassified aviation students tested in August 1942 at Psychological Research Unit No. 3, and for those of this group who entered primary pilot training in class 43E.

(1) *Distribution statistics.*—The papers of those who eventually went to primary pilot school (N=78) yielded a mean score of 7.3 and a standard deviation of 3.7 on part I (Tool Functions), and a mean score of 8.5 and a standard deviation of 3.6 on part II. At the time this test was given, there was little selection of students sent to pilot training.

(2) *Difficulty.*—The test is of approximately average difficulty. Part I yielded a mean difficulty index, corrected for chance, of approximately 0.50, while for part II, the mean difficulty, corrected for chance, was approximately 0.55, based on the 78 cases in primary training.

(3) *Factorial composition.*—Based on 153 cases, intercorrelations involving the two parts of this test were factor analyzed. Part I (Tool Functions) proved to be principally informational, having a loading of 0.77 in the mechanical-experience factor (which is best identified by the Mechanical Information test). A moderate loading (0.30) in the perceptual-speed factor and a slight loading (0.18) in the spatial-relations factor complete the list of significant factorial components of this part. The communality for Tool Functions is 0.74, which probably approaches its reliability fairly closely.

Part II has moderate loadings in mechanical experience (0.42) and in perceptual speed (0.35), and lesser loadings in the verbal (0.24), and general-reasoning (0.22) factors. The communality is only 0.41, which is probably far short of its reliability.

(4) *Test validity.*—Validation data on a small sample (N=78) of primary pilot trainees revealed unexpected results. Part I yielded only moderate validity (r_{bis}=0.17), but part II yielded a validity high enough (r_{bis}=0.40) to suggest the advisability of revising and revalidating this part of the test. Subsequent results indicated, however, that this sample was atypical, since validation of a revised form yielded a biserial of only 0.26 on a sample of 877 pilots. No item validities were computed for this test.

Evaluation.—Neither part of this test yielded stable pilot-validity figures above the middle 0.20's. Both parts are also highly correlated with more valid tests which appeared in the classification battery. For these reasons, neither part was included at any time in the classification battery.

The pilot validity to be expected for the Tool Functions test is 0.31, however, and the test's relative purity (and high loading in the mechanical-experience factor) attracts favorable attention to it. In preparing a tool-functions section, the perceptual component should be minimized by using perceptually clear and simple diagrams and by allowing liberal working time.

The pilot validity expected from the Mechanical Functions test is 0.15, based on known factors and their loadings. The indications, therefore, are that there may be an unknown factor with pilot validity in this test and that further study of it is called for. The difference between 0.15 and 0.29 is too large to be ignored, since the validity of 0.29 was obtained from a composite of more than 900 cases.

Variations of the test.—Several preliminary and subsequent forms were constructed in the course of the exploration in this area.

(1) Mechanical Operations (no code)[a]**.**—This is the original form of part II of the Mechanical Functions Test. It contains 37 scored items of the type described under part II of Mechanical Functions, CI907AX. The test was given experimentally at Psychological Research Unit No. 3 to 320 unclassified aviation students on July 9, 1942, for item-analysis purposes. The items are moderately easy, the mean difficulty, corrected for chance, being 0.58. The test is quite homogeneous, as indicated by a mean internal-consistency phi of 0.40. The best 15 items, as judged by difficulty, discriminating value of alternatives, and internal consistency, were selected to go into Mechanical Functions Test, CI907AX.

(2) Tool Function (original form; no code).—This test was also first devised as a separate test, then combined with mechanical-functions items in the Mechanical Functions test (CI907AX), and later again divorced as a separate test. The original form contains 39 scored items of the type described under part I of Mechanical Functions, CI907AX. This form was given for experimental purposes in July 1942 at Psychological Research Unit No. 3 to 360 unclassified aviation students, and an item analysis was made. It proved to be easy, the mean difficulty index, corrected for chance, being 0.61. The mean internal-consistency phi was high (0.43) for the experimental form. Items selected from this form on the basis of difficulty and internal consistency were used to form the 15-item part I of Mechanical Functions, CI907AX.

(3) Tool Function, CI906A.—This is merely a separate presentation of part I, Mechanical Functions, CI907AX, previously described.

(4) Mechanical Functions, CI907A.—This form is a separate presentation of part II, Mechanical Functions, CI907AX, with some very slight changes in wording of alternatives and in arrangement of items.

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter, Jr., T/Sgt. Paul C. Davis, Lt. Linn Hutchinson.

(5) ***Mechanical Functions, CI907B*** [8].—As a result of the promising validation data obtained for part II of Mechanical Functions, CI907AX, it was deemed advisable to revise the test and prepare a form which, if the validity held up, might be used in classification. Twelve items from part II of CI907AX were used, and 22 additional items of similar type were added. This form was administered to classified pilots in class 44C. It proved to be easier than part II of the AX form, the mean difficulty, corrected for chance, being 0.60. The items are quite homogeneous, the mean internal-consistency phi being 0.47. Validation of this form of the test on a sample of 877 pilots in primary training yielded a corrected biserial of 0.26. The elimination rate for this sample was 10 percent.

Item validation of this form showed a mean phi of 0.08.

Mechanical Movements, CI904AX2 [9]

With this test, an attempt was made to measure the assumed specific ability to comprehend and follow the operation of moving parts of machines. This test is similar to the mechanical movements test used by Thurstone in his analysis of primary mental abilities (5).

Description.—The test consists principally of questions about the movement of parts of machines. The parts are pictured, and the items, with multiple-choice answers or completing clauses, appear beside or below the drawings. Arrows and letters appear at appropriate places on the drawings to indicate parts or directions of movement. Correct answering of an item requires understanding the interaction of the parts,

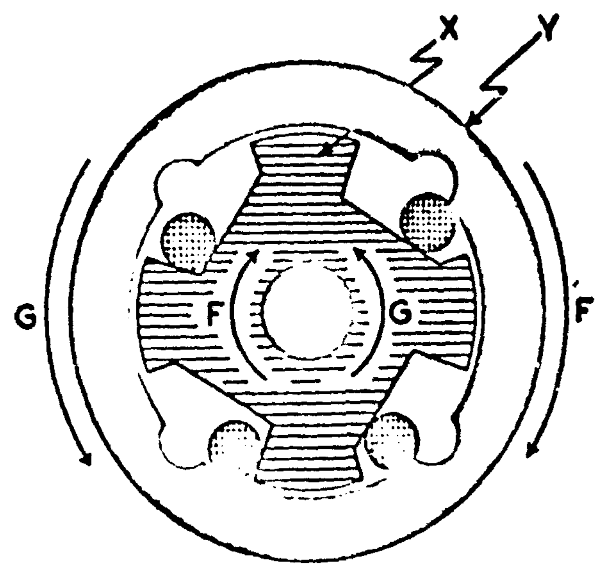


FIGURE 13.5
SAMPLE PROBLEM OF MECHANICAL MOVEMENTS, CI904AX2

[8] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, S/Sgt. Benjamin Fruchter.
[9] Developed at Psychological Research Unit No. 3. Chief contributor: T/Sgt. Paul C. Davis.

which vary in number from 2 to 12. The sample item shown in figure 13.5 is typical. The examinee is required in this problem, to select the correct alternative: "When X turns in (A) direction G, Y turns in direction G; (B) direction G, Y turns in direction F; (C) direction F, Y turns in direction F; and (D) direction F, Y turns in direction G."

(1) *Internal characteristics.*—The test consists of 2 practice items and 48 scored items. The number of alternative responses ranges from three to five per item, with an average of four. For the purpose of determining reliability, the test is separated into two equal, independently-timed parts.

(2) *Administration.*—Reading of the directions requires approximately 5 minutes. The sample diagram is accompanied by two problems followed by the correct answers. Twenty minutes are allowed for each part of the test proper. In one sample, approximately 60 percent of the examinees completed part I, while only about 35 percent finished part II. A larger proportion completed the next-to-the-last item in each part, 67 percent and 46 percent respectively, which indicates that the time is somewhat shorter than adequate to allow most to finish.

(3) *Scoring.*—The two parts of the test were scored separately, using the formula R−W/4. The part scores were summed to give a total score for use in validation and in correlating the test with other tests.

Statistical results.—This test was explored quite fully statistically. The data given below are for examinees at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Table 13.13 gives distribution data for this test.

TABLE 13.13.—*Distribution of scores for Mechanical Movements, CI904AX2*

Group	Score	N	M	SD
Unclassified aviation students[1]	R−W/4	479	22.1	7.6
Classified pilots[2]	R−W/4	674	23.0	7.4
Classified pilots[3]	R	272	22.2	5.6
Classified pilots[3]	W	272	14.0	5.4

[1] Tested February 1943.
[2] In class 43K.
[3] In class 44J.

(2) *Internal consistency.*—Item analysis of the test revealed a wide range of internal consistency. The phi values had a mean of 0.35 and standard deviation of 0.12. These data are based on the upper 25 percent and the lower 25 percent of 400 unclassified aviation students tested in April 1943.

Some consideration was given to the possibility of using some of the mechanical-movements items in a revision of the Mechanical Principles test. In order to determine which items were most highly correlated with Mechanical Principles, mechanical-movements papers of the highest and lowest 25 percent groups as determined by scores on Mechanical Principles (CI903A) were analyzed. Phi values from this analysis had a lower mean (0.27), but comparison of the two analyses revealed no significant differences in the rank order of discrimination. None of the mechanical-movements items was as closely related to Mechanical Principles total score as to Mechanical Movements total score.

(3) *Reliability coefficient.*—Reliability of the test was estimated by correlating the two part scores. This yielded a reliability coefficient (corrected) of 0.76, based on an N of 479 unclassified aviation students tested in February 1943.

(4) *Difficulty.*—Item difficulties, corrected for chance, ranged from approximately 0.00 to about 0.82 with a mean of 0.52, and a standard deviation of 0.19 based on approximately 400 cases tested in April 1943.

(5) *Factorial composition.*—This form of the test was not included in any factor-analysis study, so no data as to its factorial composition are available. Factorial composition of an earlier form and comparison with Thurstone's findings are covered in the discussion of Form A of this test.

(6) *Test validity.*—On a sample of 674 primary pilot trainees in class 43H the mean score for the graduates was 23.5 and for the eliminees 20.8. The standard deviation was 7.4, the proportion of graduates was 0.83, and the biserial correlation, 0.20.

Evaluation.—This test was not used in classification for three important reasons. In the first place, its validity is lower than validities of tests ordinarily used in classification. Secondly, it correlates highly (0.67) with the composite pilot-aptitude score. Thirdly, it also correlates highly (0.63) with the Mechanical Principles test, which yields a higher pilot validity than Mechanical Movements.

It is probable that this test might be useful as a selective device in certain mechanical areas where Mechanical Principles or a similar test would be unsuitable. Improvement of the test and investigation of its applicability to other areas might prove productive.

Variations of the test.—Several forms of this test were constructed prior to the form already described.

***(1) Mechanical Movements, CI904X1*[a]**.—This is the first form of the test, containing 58 items. It was given experimentally for the purpose of correcting and improving the items and selecting the most suitable ones for a new form of the test. Although the mean difficulty, corrected for chance, is about 0.50, many of the items are exceedingly easy, and extensive revision of others was required. This form was administered to about 400 unclassified aviation students for item analysis only in July 1942 at Psychological Research Unit No. 3.

***(2) Mechanical Movements, CI904AX*[b]**.—This form is somewhat harder than form X1, the mean item difficulty, corrected for chance, being

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter Jr., T/Sgt. Paul C. Davis, Lt. Linn Hutchinson.
[b] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter Jr., T/Sgt. Paul C. Davis, Lt. Linn Hutchinson, Lt. Llewelyn N. Wiley.

about 0.40. There is considerable range of difficulty among the 38 scored items, approximately 20 percent answering the hardest item correctly and 87 percent answering the easiest item correctly. The test was administered to several squadrons of unclassified aviation students at Psychological Research Unit No. 3 and some validation data obtained. One sample, of which 353 eventually went to primary pilot training, in class 43D, yielded a biserial of 0.21 (p_s=0.77, M_s=14.52, M_t=12.09, and SD_t=6.74). Results of an item analysis of this form were utilized in selecting items for the next form.

(3) Mechanical Movements, CI904A[a].—This form contains 5 practice items and 30 scored items selected from the experimental forms, previously referred to, on the basis of difficulty, strength of misleads, and internal consistency. The test was administered, along with several other mechanical tests, to one squadron of unclassified aviation students.

Factor analysis revealed that Mechanical Movements is extremely complicated factorially. Table 13.14 shows the principal factor loadings of this test, the loadings of the Mechanical Principles test in the same factors, Thurstone's results on his Mechanical Movements test, as reported in "Primary Mental Abilities" (5), and results of a reanalysis of Thurstone's intercorrelations.

Several important differences exist in these analyses; notably in the reasoning, mechanical, and visualization aspects. Thurstone's three reasoning factors were not well differentiated, and reanalysis reduced the number to two. Analysis of the mechanical battery revealed no deductive-reasoning factor and left the general-reasoning factor with a rather weak loading. It appears possible that what Thurstone named reasoning may have been better defined in the Psychological Research Unit No. 3 reanalysis and in the analysis of the mechanical battery as visualization, although the two would at best be only roughly equivalent. The important residuals in Thurstone's analysis probably account for much of the variance of Mechanical Movements. Much of this residual variance might well be the mechanical factor that appears in the analysis of the Psychological Research Unit No. 3 Mechanical Battery. Since no other mechanical tests appeared in Thurstone's battery, no common mechanical factor could be defined.

As indicated by table 13.14, Mechanical Movements and Mechanical principles have much in common factorially. The only important exceptions are the greater loading in the perceptual-speed factor, and the much smaller loading in the mechanical-experience factor for Mechanical Movements. The much higher pilot validity (0.34) of Mechanical Principles than of Mechanical Movements (0.23) is largely due to the fact that the mechanical-experience factor is a more important determiner of pilot validity than is perceptual speed.

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter Jr., T/Sgt. Paul C. Davis, Lt. Lann Hutchinson, Lt. Llewelyn N. Wiley.

TABLE 13.14.—*Results of various factor analyses, showing composition of Mechanical Movements tests and Mechanical Principles tests*

Analysis	Test	Factors						
		Spatial relations	Visualization	Perceptual	Mechanical experience	Reasoning (inductive)	Reasoning (deductive)	Reasoning (general)
Thurstone's analysis	Mechanical Movements.	0.07	....[1]	0.19	....	0.17	0.40	0.41
Psychological Research Unit No. 3 analysis of Thurstone data	Mechanical Movements.	.15	.32	.17	....	....	.56	.45
Psychological Research Unit No. 3 analysis of mechanical battery	Mechanical Movements.	.28	.51	.37	0.38	....	....	.25
Psychological Research Unit No. 3 analyses	Mechanical Principles	.22	.51	.08	.60	....	....	.20

[1] The symbol indicates that the factor did not appear in the analysis.

Summary and Evaluation of Mechanical Comprehension Tests

As indicated by the statistical data for the tests in this section, positive results were attained in the search for abilities correlated with success in air-crew training, especially that of navigator and pilot. Outstanding is the relatively high validity of the Mechanical Principles test, two forms of which were in classification batteries during most of the period covered by this report. This validity results principally from its loadings with the mechanical-experience and visualization factors (especially for the pilot) and to some degree from its loading with the spatial-relations factor. The latter factor also contributes a considerable proportion of the navigator validity of the test and a small amount of the limited bombardier validity.

Although validities against air-crew training criteria are lower, in general, for other tests in this section than for Mechanical Principles, sufficiently high validities were obtained for all tests to indicate their potential usefulness as selection devices for other mechanical tasks.

Examination of the factorial composition of the tests in this section reveals a great deal of similarity among them. The variable most prominent in all these tests and the one probably largely responsible for their validities is one apparently best defined as mechanical experience. Indications are, however, that differences in validity among the tests of this group are due to variations in the total factorial picture rather than to the loadings in any one factor. The fact that validities of some magnitude were found for most other factors appearing in these tests, namely, the perceptual-speed factor, the spatial-relations factor, and the visualization factor, supports this view. Determination of the extent to which tests of such complicated factorial composition as these are more generally applicable in the selection of individuals for mechanical tasks must depend upon future research. It is not known whether all mechanical jobs stress this particular combination of fundamental abilities.

MECHANICAL INFORMATION

Definition and Rationale

In this group are included tests that consist principally of informational items in the field of mechanics. Most of the material is presented verbally, although the Driving Skill and Physical Principles tests utilize some pictorial material. In general, the correct answers to items in these tests cannot be determined by reasoning but require specific knowledge. This is one respect in which the tests in this group differ from those in the mechanical-comprehension section.

The objective in these tests is to evaluate some aspects of mechanical ability by measures of information. Although it was recognized that the possession of mechanical information alone does not constitute qualification, per se, for mechanical tasks such as those of air-crew members, it appeared likely that possession of such information would be symptomatic of the presence of certain other characteristics essential to the successful performance of mechanical tasks. In line with this view, it was not considered necessary to construct the tests in this section with any particular reference to air-crew functions.

Mechanical Information, CI905A [11]

The items in this test cover information about the structure, function, disfunction, and repair of machines. Major emphasis is placed upon automotive information, 26 of the 30 items being related to automobiles. Approximately one-half of the items are very brief, as illustrated by the following sample item:

A fuel pump is driven by the:

A. Flywheel
B. Fan belt.
C. Generator shaft.
D. Cam shaft.
E. Distributor shaft.

The other items are much longer and cover descriptions of situations involving mechanical problems. The following item is illustrative of the latter type of item in the test:

With pressure on the starter, the starting motor runs smoothly, but no contact is made between the starting motor and the engine. The most probable cause of the trouble is that:

A. The armature of the starting motor is loose.
B. The brushes in the starting motor are not making contact with the commutator.
C. The bendix spring is broken.
D. The fuse is blown.
E. The ignition coil is not functioning properly.

(1) *Administration.*—The time allowed is 15 minutes. Approximately 85 percent of the group (unclassified aviation students) finishes the test in the time allotted.

(2) *Scoring.*—The scoring formula is R—W/4.

Statistical results.—Experimental and classification-battery administration of this test yielded numerous statistical data.

(1) *Distribution statistics.*—Administered to more than 3,000 aviation students, the test yielded the distributions given in table 13.15. The distribution curves are approximately symmetrical, but with greater than normal frequencies in the extreme upper and lower reaches. It is interesting to note that the almost identical B form yielded a mean of 12.3 and a standard deviation of 7.4 for West Point cadets, using R—W/3 in place of the customary formula, R — W/4.

[11] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, Lt. Linn Hutchinson.

TABLE 13.15.—*Distribution of scores of unclassified aviation students on Mechanical Information, CI905A*

N	M	SD
[1]1,096	15.3	8.3
[2]1,015	14.1	8.5
[3]1,143	14.5	8.3

[1] Tested at Psychological Research Unit No. 1 with the December 1942 Classification Battery.
[2] Tested at Psychological Research Unit No. 2 with the December 1942 Classification Battery.
[3] Tested at Psychological Research Unit No. 3 with the December 1942 Classification Battery.

(2) *Internal Consistency*—Item analysis of the test, based on the upper 25 percent and the lower 25 percent of 360 unclassified aviation students tested at Psychological Research Unit No. 3, revealed a high degree of homogeneity, as indicated by a mean phi of 0.56, a standard deviation of 0.09, and a range from 0.05 to 0.80. From these data it may be seen that the test is one of the most homogeneous devised for air-crew classification.

(3) *Reliability coefficient.*—Since most of the examinees finish within the time limit, the odd-even method of estimating reliability was employed. Based upon two different samples of 240 cases each, tested in April 1943 at Psychological Research Unit No. 3, an average corrected reliability of 0.89 was obtained, the two figures being 0.88 and 0.90.

(4) *Difficulty.*—The mean difficulty index of the items, corrected for chance, is 0.48, with a range from 0.24 to 0.71 and a standard deviation of 0.14, based on the above-mentioned sample of 360 unclassified aviation students.

(5) *Factorial composition.*—This test proved to be one of the purest measures of any single factor. Since it appeared in several analyses, the factorial composition can be considered to be quite reliably ascertained. The only significant loading (0.74) is in the factor identified as mechanical experience. Slight loadings in visualization (0.15) and verbal (0.11) factors are found, but these factors contribute very little to the test.

(6) *Test validity.*—Validation data are presented in tables 13.16 through 13.18.

(7) *Item validity.*—The mean phi was 0.047, with a standard deviation of 0.046 and a range from −0.02 to 0.10, based upon a sample of 800 graduates and 600 eliminees from primary pilot training in classes 43K through 44C, originally tested at Psychological Research Unit No. 3.

Evaluation.—From these data it may be concluded that this test has extensive possibilities, although some of the evidence is inconclusive. It appears that this test is most useful in predicting success in tasks in which mechanical experience is highly important. Pilot training, air-mechanic grades, radio-operator-mechanic average grades and flexible-gunnery final-examination grades are in this class. Why academic averages for armorers and flexible gunners are not more highly correlated with scores on this test is not entirely clear. It seems probable, however, that such academic grades are more heavily weighted with non-mechani-

TABLE 13.16.—*Validation data for Mechanical Information, CI905A, based upon graduation-elimination criteria*

Group	N	Class	Psychological research unit No.	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
Pilots in primary training	513	43G	3	0.75	17.2	13.9	8.2	0.24	....
Pilots in primary training	652	43H	2	.59	15.6	11.2	8.5	.32	....
Pilots in primary training	3,150	43I	1, 2, 3	.66	15.9	12.9	8.1	.23	0.28
Pilots through advanced training	2,978	43I	1, 2, 3	.54	16.1	13.2	8.1	.22	.27
Pilots in B-17 transition training	1,046	43I, K	1, 2, 3	.98	15.9	11.8	7.9	.27	....
Pilots in B-24 transition training	982	43I, K	1, 2, 3	.92	16.1	13.7	7.9	.15	....
Pilots in B-25 transition training	314	43I, K	1, 2, 3	.98	16.2	18.7	8.0	−.1[illegible]	....
Pilots in B-26 transition training	380	43I, K	1, 2, 3	.82	16.0	16.0	8.0	.00	....
Pilots in P-40 assignment	2,415	43I, K	1, 2, 3	.64	14.8	14.7	8.0	.01	....
Pilots in primary training	1,951	43I	3	.74	[2]5.4	4.8	1.9	.18	....
Pilots in primary training	2,497	43K	3	.81	[2]5.6	5.2	1.9	.12	....
Navigators	731	43-10, 11	1, 2	.87	11.0	9.5	8.2	.10	.14
Navigators	1,970	43-12 through 43-15	1, 2, 3	.79	11.3	11.3	8.3	.00	.05
Bombardiers[3]	1,828	43-8 through 43-11	1, 2, 3	.79	10.2	10.5	7.7	−.02	.00
Bombardiers[4]	455	43-14 through 43-18	1, 2, 3	.84	9.4	8.2	7.5	.09	.10
Bombardiers[5]	524	43-14 through 43-18	1, 2, 3	.86	14.4	12.3	8.0	.14	.17

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Using stanine scores with a mean of 5.00 and a standard deviation of 2.00.
[3] New aviation cadets taking the 12-week course (no navigation training).
[4] New aviation cadets taking the 18-week course (some navigation training).
[5] Reclassified pilots taking the 18-week course.

703320—47—22

TABLE 13.17.—*Validation data for Mechanical Information, CI905A, for seven grades in navigation training*[1]

Criterion	r	$r_{corr.}$
Grades in dead reckoning (ground school)	0.04	0.08
Grades in celestial navigation (ground school)	−.03	.02
Grades in dead reckoning (flight)	.02	.04
Grades in celestial navigation (flight)	.03	.06
Grades in meteorology	.05	.09
Military grades	−.02	.00
Final composite grades	.01	.06

[1] For a sample of 463 navigation trainees in Hondo classes 43-10 through 43-15, tested at Psychological Research Units Nos. 1, 2, and 3.

TABLE 13.18.—*Validities of Mechanical Information, CI905A, for technical specialties*

Specialty	Criterion	N	r[1]
Air mechanic[2]	Average grades	232	0.49
Armorer[2]	Average grades	376	.19
Radio operator-mechanic[2]	Average grades	153	.33
Flexible gunnery[3]	Air-to-air[4]	61	.00
Flexible gunnery[5]	Air-to-air[4]	194	−.10
Flexible gunnery[6]	Air-to-air[4]	173	.15
Flexible gunnery[3]	Academic average	61	.00
Flexible gunnery[5]	Final examination	194	.31
Flexible gunnery[6]	Final examination	173	.36

[1] Product-moment correlations.
[2] Tested at Psychological Research Units Nos. 1, 2, and 3 with the December 1942 Battery.
[3] In classes 43-27 to 43-30 at Tyndall. Tested at Psychological Research Unit No. 1.
[4] A very unreliable criterion.
[5] In class 43-45 at Ft. Myers. Tested at Psychological Research Units Nos. 1, 2, and 3.
[6] In class 43-48 at Ft. Myers. Tested at Psychological Research Units Nos. 1, 2, and 3.

cal material than are the tasks for which higher validities were obtained. This test is a relatively pure measure of the mechanical-experience factor and would probably be useful in predicting success in most other mechanical tasks.

Variations of the test.—Considerable attention was given to this area with the result that several forms were constructed.

(1) Mechanical Knowledge (no code)[12].—This is the first form of the mechanical-information type of test. It contains 40 scored items, 27 of which are directly related to automobile or airplane engines. The test was administered in July 1942 at Psychological Research Unit No. 3 to 360 unclassified aviation students, and an item analysis was made. This form proved to be homogeneous (mean internal-consistency phi=0.47), but is somewhat too easy. The mean difficulty index, corrected for chance, is 0.58. On the basis of the item analysis, several of the items were selected for use in Form A of mechanical information, already described.

(2) Mechanical Information, CI905AX2[13].—This is the original form by this name. It contains 25 items about automobiles, including diagnosis of trouble (20 items) and functioning of the machine and its parts (5 items). This form is moderately difficult (mean difficulty index, corrected

[12] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, Lt. Linn Hutchinson.
[13] Same as footnote 12.

for chance, = 0.45 for an N of 350 students tested at Psychological Research Unit No. 3). The test is quite homogeneous, yielding a mean internal-consistency phi of 0.41 with a standard deviation of 0.12, for the above-mentioned sample. The number of items in this form was considered inadequate in view of the revisions expected. The item-analysis data were helpful in preparing the revised form.

(3) Mechanical Information, CI905AX2[14].—This 35-item form is a revision of Form AX1 with additional items of the same type described under AX1. It was administered for item-analysis purposes to 400 unclassified aviation students in June 1942 at Psychological Research Unit No. 3. The mean phi was 0.38, with a standard deviation of 0.16 and a range from 0.09 to 0.80. The mean difficulty of the items, corrected for chance, is 0.44, and the standard deviation of difficulty values is 0.20.

Fifteen of the items from this test that yielded high internal-consistency phis and satisfactory difficulty were selected for use in Mechanical Information, CI905A.

(4) Mechanical Information, CI905BX.—This is a two-choice form of CI905A, prepared for the purpose of studying item reliability. Based upon the item analysis of CI905A, the correct answer and the mislead with the highest discriminating value were selected for use in this form. Item analysis based upon experimental administration of the test at Psychological Research Unit No. 3 in November and December 1943 to 800 unclassified aviation students yielded a mean internal-consistency phi of 0.47 and a standard deviation of 0.08.

Although the mean phi values for this form are lower than for Form A, these data show that relatively high internal consistency (hence reliability) can be achieved even with two-alternative items, if misleads are strong discriminators between good and poor groups as determined by total score on the test.

(5) Mechanical Information, CI905B.—This is a second classificaiton form of mechanical information and differs from CI905A only slightly. Some alternatives were revised, omitted, or rearranged, but otherwise little change was made in preparing this form.

Driving Skill, CI307AX1[15]

Because of the similarity, superficial, at least, of the tasks involved in driving an automobile or truck and in flying an airplane, it appeared logical to expect that a measure of success in the former would constitute a good predictor of success in the latter. It could be assumed that experience with, and hence knowledge of, the operation of automobiles might be indicative of interest in and certain aptitude for mechanical tasks.

[14] Same as footnote 12.
[15] Developed at Psychological Research Unit No. 3. Chief contributors: T/Sgt. Paul C. Davis, Major James J. Gibson, Capt. L. G. Humphreys, and Lt. Linn Hutchinson. An earlier form, called "Automobile Driving Test," was developed by Major Neal E. Miller.

It was originally reasoned that a test of driving skill might measure a type of judgment that is important in flying. On the basis of this assumption, this test was given a Judgment code number. As will be seen later in the discussion, this assumption did not prove to be correct.

Description.—The test consists of 42 scored items, 31 presented verbally, and 11 presented by means of pictures, representing situations in which the examinee is required to indicate the best decision or driving practice. The following item and figure 13.6 are typical of the two kinds of items.

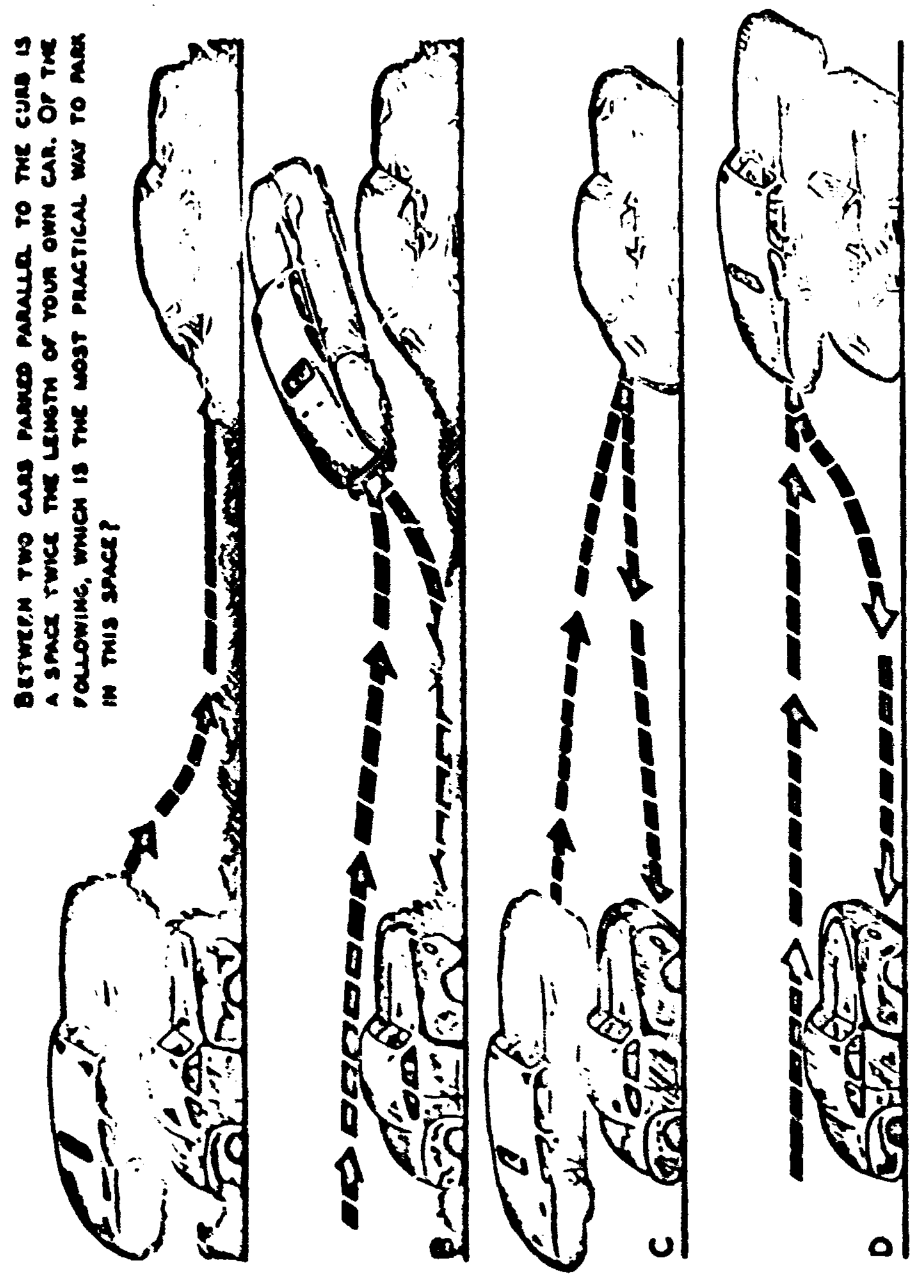


FIGURE 13.6
SAMPLE PROBLEM OF DRIVING SKILL, CI307AXI

If one front tire is softer than the other, the car will tend to

A. Pull to the side of the soft tire.
B. Skid in the direction of the soft tire.
C. Pull away from the side of the soft tire.
D. Skid in the opposite direction from the soft tire.
E. Pull from one side of the road to the other.

(1) *Internal characteristics.*—The test is separated into two parts of 21 items each. Each part is timed separately (15 minutes) in order to provide a basis for estimating reliability.

(2) *Administration.*—Because of the unusual problems in this test, it was necessary to include a special paragraph in the directions. This paragraph warns against answering according to legal rules and urges the examinee to answer in line with best driving practice without reference to traffic regulations. Standard directions for answering and marking answer sheets are also included. The time allowed permits approximately 80 percent of the students to finish the test. This is quite satisfactory, inasmuch as the test is designed as a power test.

(3) *Scoring.*—Most of the items contain five alternative responses, but a few have only three or four. The test is scored R—W/4.

Statistical results.—This test appeared in a battery of tests which underwent rather thorough statistical analysis. Except where noted to the contrary the data are based upon examinees tested in December 1942 at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—A sample of 202 unclassified aviation students yielded a mean score of 18.6 and a standard deviation of 5.4. The distribution of scores is symmetrical, but slightly flatter than normal.

(2) *Internal consistency.*—An internal-consistency item analysis of the test, based on the upper 25 percent and the lower 25 percent of 202 unclassified aviation students, yielded a mean phi of 0.28, a standard deviation of 0.11, and a range from 0.08 to 0.58. Twelve items have phi values below 0.20. Examination of these items reveals two probable reasons for their lack of internal consistency. In some instances certain alternatives are general and cover other more specific alternatives also, leaving the best (not to say correct) answer questionable. In many cases, also, the correct answer is poorly presented or answering is dependent upon correct interpretation of the situation. Almost half (five) of the pictorial items are in the low-phi group.

(3) *Reliability coefficient.*—By the alternate-forms method (pt. I-pt II), the reliability is 0.55, corrected for length, based upon a sample of 239 unclassified aviation students.

(4) *Difficulty.*—The difficulty levels of the items based on the item analysis previously referred to is indicated by a mean proportion of correct responses, corrected for chance, of 0.47, a standard deviation of 0.21, and a range from 0.00 to 0.85.

(5) *Factorial composition.*—The loading in the mechanical-experience factor was found to be 0.46 and that in visualization 0.42. A small

loading (0.15) also is found in spatial relations. The test appears to measure about the same things as the mechanical tests (Mechanical Principles, CI903A, and Mechanical Movements, CI904A), but much less reliably. Its communality of 0.53 almost completely exhausts its nonerror variance.

(6) *Test validity.*—Based upon a relatively small sample of 149 pilots in primary training in class 43J, the test yielded a biserial validity of 0.12. The mean of the graduates was 20.34, that of the eliminees 19.34, and the standard deviation of the total, 4.92. Of this sample, 75 percent were graduates.

(7) *Item validity.*—A study of item validity for pilots, based upon 200 graduates and 45 eliminees from primary training, showed validity phis ranging from −0.13 to 0.28 with a mean of 0.03 and a standard deviation of 0.09. This sample was tested in November 1943.

Evaluation.—This test was included in the first foresight-and-planning battery, being considered at that time to be a measure of planning and judgment. Factor analysis of the intercorrelations revealed that, contrary to the original hypothesis, the driving skill test is most heavily loaded in the mechanical-experience and visualization factors rather than in planning.

From its factors and their loadings one would expect a pilot validity of 0.30 for this test. The obtained composite validity of 0.32 (see table 28.18) is very close to this expectation.

Its factorial content indicates that any use to which it might be adapted could be better performed by other mechanical tests. Certain items that correlate highly with a mechanical-information test score and low with a visualization test score, might well be incorporated in a test of the mechanical-experience factor. Others that prove to be valid for pilots might be incorporated in a heterogeneous general-information test. In view of the apparent duplication in this test of the functions measured by other mechanical tests, it was not deemed profitable to develop this test, as such, further for pilot selection.

Physical Principles, CI801BX [16]

The dependence of all mechanical phenomena upon basic laws of physics was viewed as being sufficient justification for a test of technical physics, at least experimentally. It was reasoned that if mechanical ability should prove valid for the prediction of air-crew success, knowledge of the basic principles upon which mechanics depend might also be valid. It also appeared that knowledge of the correlations of mechanical tests with a test of formal physics would be valuable in analyzing the results and evaluating the usefulness of such tests. One fact tending to

[16] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. John A. Bath, Lt. Mahlon B. Smith.

indicate that a test of formal physics might not be valid, at least for pilot selection, was the approximately zero correlation of academic intelligence (verbal ability) with pilot success in primary training. All in all, these considerations suggested that a measure of knowledge of academic physics should be constructed and the results of its application thoroughly examined.

Description.—This test contains 30 items, 28 of which are presented verbally and 2 diagrammatically. The items are based predominantly upon principles and laws of physics which would be fairly familiar to a student who has just completed a course in high-school physics. The electricity items are generally at a relatively simple level. Previous forms of the test had provided indications that electricity items of greater difficulty were so hard as to be practically useless in discriminating between good and poor groups, since few were able to answer them.

(1) *Administration.*—The time allowed is 18 minutes.

(2) *Scoring.*—The test is scored R−W/4.

Statistical results.—Fuller statistical results are available for this than for other forms of the test. The data are for examinees tested at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Distribution data are given in table 13.19.

TABLE 13.19.—*Distribution data for Physical Principles, CI801BX*

Group	N	M	SD
Unclassified aviation students[1]	368	11.9	7.6
Classified pilots[2]	5,408	16.8	7.1

[1] Tested in October and November 1942.
[2] In classes 44B and 44C. Tested in November 1943.

(2) *Internal consistency.*—As is true for most of the tests in this area, the items of this test proved to be highly cohesive. The mean internal-consistency phi based upon the upper 25 percent and lower 25 percent of 800 unclassified aviation students tested in November 1943 is 0.51, and the standard deviation is 0.09. The range of values is relatively small (0.32 to 0.66).

(3) *Reliability coefficient.*—In view of the fact that the test is essentially a power test, the odd-even method of estimating reliability was employed. Based on the above-mentioned sample of 368 (see table 13.19), the estimated reliability, corrected for length, is 0.86.

(4) *Difficulty.*—The mean difficulty index of the items, based on the sample referred to previously (N=368), is 0.44, corrected for chance. The standard deviation of the distribution of corrected difficulties is 0.15. The range of corrected difficulties (0.14 to 0.72) indicates adequate variety of difficulty for the group being tested.

(5) *Factorial composition.*—The highest loadings of the test are in the mechanical-experience (0.51) and verbal (0.38) factors. No others exceeded 0.20. The communality of 0.46 is far below the test reliability.

(6) *Test validity.*—This test was validated against the primary graduation-elimination criterion in a sample of 5,408 pilots in classes 44B and 44C, of whom 93 percent graduated. Graduates had a mean score of 18.96, and eliminees a mean of 17.40. The standard deviation of the total was 7.08, and the biserial correlation was 0.15, corrected to an assumed unrestricted stanine standard deviation of 1.81.

Evaluation.—This test measures mechanical experience but less purely than several other tests, particularly the mechanical-information tests. Its obtained validity (see Table 28.18) fell slightly short of its expected validity of 0.15 for pilot selection.

A test such as this might prove very useful in selecting and classifying for tasks in which both academic intelligence and mechanical knowledge play an important part. The homogeneity and the substantial verbal loading suggest its possible use in predicting success in technical or engineering studies.

Variations of the test.—Several other forms of physics tests were prepared.

(1) Physics, CI801A[11].—This is a preliminary 30-item form that was constructed for administration in August 1942 at Psychological Research Unit No. 3 with the experimental mechanical battery. The test consists of verbally-presented technical physics items and was administered to about 250 unclassified aviation students. It proved to be too hard, the mean difficulty index, corrected for chance, being approximately 0.34, and the standard deviation being 0.19. Factorially, this form bears close similarity to Reading Comprehension, CI614H, described in chapter 5, its loadings in the verbal (0.68), visualization (0.25), and reasoning (0.17) factors differing only slightly from those of Reading Comprehension in the same factors (0.60, 0.30, and 0.19 respectively). The loading in mechanical experience is only 0.21, as compared with 0.51 in the B form.

(2) Physical Principles, CI801AX.—This form contains a total of 102 items and constitutes an attempt to bring together a large number of all types of physics problems, from which a shorter form could be prepared. This form proved to be of approximately the same difficulty as the 30-item preliminary form, the mean difficulty, corrected for chance, being 0.35 and standard deviation being 0.19. This form is quite homogeneous, yielding a mean internal-consistency phi of 0.43 and a standard deviation of 0.13, based upon 360 cases tested in August 1942 at Psychological Research Unit No. 3. Selection of items for the CI801BX form was made

[11] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter Jr., T/Sgt. Paul C. Davis, Lt. Linn Hutchinson, Lt. Llewelyn N. Wiley.

on the basis of (1) suitable difficulty, (2) availability of functional misleads, and (3) appropriate internal-consistency phi values.

Summary and Evaluation of Mechanical Information Tests

The tests included in this group have one outstanding characteristic in common; they call for knowledge of specific facts. The specificity of these facts resulted in quite diverse factorial content among the various tests. The outstanding example is the relatively high loading in the mechanical-experience factor achieved by the Mechanical Information and Driving Skill tests, while the Physics test, on the whole, was most heavily loaded in the verbal factor. The relatively high pilot validity of the mechanical-experience factor made the tests high in this factor most useful in pilot selection. The homogeneity of most of the tests in this group suggests that each of the tests might be used to good advantage in situations where specificity of function, such as that involved in the test, is required. These tests would probably be less useful for general predictive purposes than those included in the mechanical-comprehension section.

A FACTOR ANALYSIS OF MECHANICAL TESTS

Although mechanical tests have been in use for some time in civilian life, as indicated in the introduction to this chapter, little was known about their unique components or factors. After the preparation of a battery of mechanical tests, it appeared desirable to factor analyze intercorrelations among these and certain other tests. These other tests were added in order to obtain a comprehensive picture of the mechanical tests by bringing out as many of the factors as possible.

The Data

This analysis included 17 tests which covered many phases of human ability. Of these tests, seven were designed as strictly mechanical tests, three involve length estimation (Nearest Point, Shortest Path, Shorter Line), and three were designed along with mechanical tests but were later found to be quite different (Physics, Pattern Assembly, Pattern Comprehension). The remaining four (Reading Comprehension, Arithmetic Reasoning, Spatial Orientation, Complex Coordination) represented different areas covered by the classification tests. All these tests, with the exception of Complex Coordination—an apparatus test—and Mechanical Comprehension, AC10B and AC10D, are described elsewhere in this volume. The last-named tests are, superficially, at least, quite similar to the Mechanical Functions test. They were parts of the two AAF qualifying examinations with the code numbers given (see report No. 6).

The list of tests and intercorrelations appears in table 13.20. The correlations are based on a sample of 153 unclassified aviation students. Table 13.21 presents centroid loadings and rotated factor loadings.

Table 13.20.—*Table of intercorrelations of the Mechanical Battery*[1]

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Mechanical Principles, CI903A	..	65	50	56	35	33	19	47	26	53	55	24	25	38	23	29	24
2. Mechanical Movements, CI904A	65	..	40	40	38	35	25	38	24	50	47	33	44	40	28	42	12
3. Mechanical Information, CI905A	56	40	..	64	37	07	04	23	10	44	43	09	08	26	19	17	13
4. Tool Function, CI906A	50	40	64	..	43	19	18	22	11	41	30	04	20	24	20	12	18
5. Mechanical Functions, CI907A	35	38	37	43	..	28	20	38	31	29	34	24	27	19	13	26	10
6. Pattern Comprehension, CP803A	33	35	07	19	28	..	22	36	27	32	37	36	35	15	27	36	04
7. Pattern Assembly, CP804A	19	25	04	18	20	32	..	16	04	12	16	25	28	14	40	34	35
8. Physics, CI801A	47	38	23	22	38	36	16	..	57	45	44	28	28	15	11	28	19
9. Reading Comprehension, AC10D	26	24	10	11	31	27	04	57	..	35	38	39	29	20	02	18	07
10. Mechanical Comprehension, AC10B	53	30	44	41	29	32	12	45	35	..	51	33	28	39	24	22	11
11. Mechanical Comprehension, AC10D	55	47	43	30	34	37	16	44	38	51	..	24	24	21	17	18	18
12. Arithmetic Reasoning, CI206B	24	33	09	04	24	25	25	28	39	33	24	..	23	22	36	30	24
13. Spatial Orientation, CP501A	25	44	08	20	27	35	28	28	29	28	24	23	..	23	09	41	21
14. SAM Complex Coordinator	38	40	26	24	19	15	14	15	20	39	21	22	23	..	30	25	23
15. Nearest Point, CP607B	23	28	19	20	13	27	40	11	02	24	17	36	09	30	..	32	37
16. Shortest Path, CP608B	29	42	17	12	26	36	34	28	18	22	18	30	41	25	32	..	31
17. Shorter Line, CP606B	24	12	13	18	10	04	35	19	07	11	18	24	21	23	37	31	..

[1] Decimal points have been omitted.

TABLE 13.21.—*Factor loadings of the Mechanical Battery*[1]

Test	Centroid loadings I	II	III	IV	V	VI	VII	h^2	Rotated loadings I	II	III	IV	V	VI	VII	h^2
1. Mechanical Principles, CI903A	72	25	15	05	14	−23	−06	682	54	12	19	06	49	29	09	678
2. Mechanical Movements, CI904A	73	06	−04	−14	27	−17	−16	685	38	37	05	−04	51	28	25	686
3. Mechanical Information, CI905A	52	44	37	−11	10	09	11	643	78	00	07	03	12	01	11	641
4. Tool Function, CI906A	55	32	41	−29	−13	24	−12	746	77	30	−02	11	−11	18	05	742
5. Mechanical Functions, CI907A	54	16	−04	−22	−16	13	07	415	42	35	24	05	04	−03	22	410
6. Pattern Comprehension, CP803A	51	−13	−32	−08	11	05	13	417	06	29	21	00	28	−01	45	413
7. Pattern Assembly, CP804A	41	−41	12	−10	−17	−09	14	417	02	31	05	52	16	01	14	415
8. Physics, CI801A	60	22	−32	17	−26	−10	10	627	21	17	68	01	25	03	17	628
9. Reading Comprehension, AC10D	48	15	−46	28	−24	13	−07	622	05	15	73	−01	02	14	22	626
10. Mechanical Comprehension, AC10B	66	22	−03	17	18	08	−10	563	43	06	25	−09	27	36	32	564
11. Mechanical Comprehension, AC10D	62	27	−08	11	08	−13	11	511	42	02	35	−01	40	12	19	510
12. Arithmetic Reasoning, CI206B	50	−26	−22	21	08	21	10	471	−04	10	25	15	14	19	56	466
13. Spatial Orientation, CP501A	50	−18	−24	−28	−11	−10	−23	493	02	62	16	02	24	13	12	500
14. SAM Complex Coordinator	47	−08	16	14	14	06	−31	392	20	14	02	.08	15	52	16	385
15. Nearest Point, CP607B	45	−40	30	14	17	16	21	571	11	−02	−12	50	16	20	47	563
16. Shortest Path, CP608B	53	−36	−08	−20	03	−12	08	479	−01	06	10	46	32	27	26	468
17. Shorter Line, CP606B	38	−34	29	20	−25	−21	09	499	04	05	15	63	20	20	−01	504

[1] Decimal points have been omitted.

The Factors

In this section the factors are described, and the tests appearing on each are listed in order of decreasing saturation. Only rotated factor loadings of 0.25 and above are given.

Rotated factor I is defined by the following data:

Test No.	Test name	Loading
3	Mechanical Information	0.78
4	Tool Function	.77
1	Mechanical Principles	.54
10	Mechanical Comprehension, AC10B	.43
11	Mechanical Comprehension, AC10D	.42
12	Mechanical Functions	.42
2	Mechanical Movements	.38

This is the mechanical-experience factor. At one time it was designated mechanical information, but this and other analyses tend to indicate that experience or background are better terms, since many tests appearing on the factor, such as Mechanical Principles and Mechanical Movements, seem to depend more upon general mechanical experience than upon specific knowledge of things mechanical. Not a single mechanical test in the battery failed to have a substantial loading on this factor. On the other hand, tests like Pattern Comprehension and Pattern Assembly, traditionally believed to be in the mechanical area, are conspicuous by their absence. If they are valid for predicting success in mechanical tasks, it is because of some other factor than this one. It is believed that those other factors have been identified in this and other analyses reported in this volume.

Rotated factor II is defined by the following data:

Test No.	Test name	Loading
13	Spatial Orientation I	0.62
2	Mechanical Movements	.37
5	Mechanical Functions	.35
7	Pattern Assembly	.31
4	Tool Function	.30
6	Pattern Comprehension	.29

This is the perceptual-speed factor defined in other analyses by Speed of Identification as well as Spatial Orientation. Two of the mechanical tests, which involve following operations of parts, have moderate loadings in this factor but neither has a high loading. Pattern Assembly (a paper form board) and Pattern Comprehension (surface development) also appear on this factor, although the loadings indicate that they are not primarily perceptual tests. The absence of Mechanical Information from this list was to be expected. The presence of the Tool Function test must mean that the diagrams in that test were too detailed, or the element of speed was somehow stressed too much, or both features share the blame for the loading on perceptual speed.[16] Since this finding has not been verified in a second analysis, however, not too much concern should be felt about explaining it or about ridding the test of it.

Rotated factor III is defined by the following data:

Test No.	Test name	Loading
9	Reading Comprehension	0.73
8	Physics	.68
11	Mechanical Comprehension, AC10D	.35
10	Mechanical Comprehension, AC10B	.25
12	Arithmetic Reasoning	.25

This is the verbal factor, which has been well defined in several analyses discussed elsewhere in this volume. The loading of 0.68 for Physics in this factor and its loading of 0.21 in mechanical experience emphasize the great factorial difference between this test and the mechanical tests in general. The two mechanical-comprehension tests appearing on this factor display more characteristics of verbal tests than do the tests described in this chapter. The absence of the Mechanical Information test from the list is most eloquent of the possibility of exclusion of undesired factors from tests. Its items are entirely verbally presented, and yet the level of verbal comprehension is apparently so low that individual differences in the trait do not influence scores in the test. The test of Driving Skill that, in another analysis, had a high mechanical variance but zero verbal variance is another good example.

Rotated factor IV is defined by the following data:

Test No.	Test name	Loading
17	Shorter Line	0.63
7	Pattern Assembly	.52
16	Nearest Point	.50
15	Shortest Path	.46

This is a length-estimation factor, which is quite well defined by the three quantitative-perception estimation tests (see ch. 18). Conscious effort had been made to introduce length estimation into the Pattern Assembly test, and its loading of 0.52 on this factor indicates that solution of the problems depends to a considerable extent on this factor. The factor should possibly be defined as a more general "size-estimation" ability, because it must be recognized that the tests listed represent different kinds of space judgments—length of lines (straight lines or irregularly curved lines), gaps between points, and gaps between lines. Thus, it would seem that not only one-dimensional but also two-dimensional extents are discriminated in the tests loaded with this factor. It is possible that the more complex discriminations rest upon combinations or abstrac-

[16] As a matter of fact, the analyzed form of this test was mimeographed, and some diagrammatic reproductions were poor where details were important.

tions of linear judgment, however. Since the test with the greatest loading is Line Length, also, we are probably justified in choosing the factor title given—length estimation.

Rotated factor V is defined by the following data:

Test No.	Test name	Loading
2	Mechanical Movements	0.51
1	Mechanical Principles	.49
11	Mechanical Comprehension, AC10D	.40
16	Shortest Path	.32
6	Pattern Comprehension	.28
10	Mechanical Comprehension, AC10B	.27
8	Physics	.25

This is the visualization factor, which also appeared in several other analyses. It seems to be a very common secondary factor in mechanical tests, particularly in those that involve moving mechanisms. This gives one small clue as to the nature of the factor which has been tentatively defined as a manipulatory visualization. Objects are imagined as moving or as having been moved or transformed in tests loaded with this factor.

Rotated factor VI is defined by the following data:

Test No.	Test name	Loading
14	SAM Complex Coordination	0.52
10	Mechanical Comprehension, AC10B	.36
1	Mechanical Principles	.29
2	Mechanical Movements	.28
16	Shortest Path	.27

This is the spatial-relations factor, long defined by the Complex Coordination test. This factor appears generally in pictorial mechanical tests but not in verbal tests, as might be expected. The spatial arrangement of mechanical devices seems to be the significant element bringing this about.

Rotated factor VII is defined by the following data:

Test No.	Test name	Loading
12	Arithmetic Reasoning	0.56
15	Nearest Point	.47
6	Pattern Comprehension	.45
10	Mechanical Comprehension, AC10B	.32
16	Shortest Path	.26
2	Mechanical Movements	.25

This appears to be the general-reasoning factor isolated in other analyses. The loadings of Arithmetic Reasoning, Pattern Comprehension, and Mechanical Comprehension, particularly, clearly agree with this naming of the factor. The loading of 0.47 for Nearest Point is very difficult to rationalize. The magnitude of the loading may be a sampling artifact, however, since the correlation of the test with Arithmetic Reasoning for a larger sample (N=392) was only 0.10 as compared with 0.36 in this matrix. The heterogeneous list of tests is not unusual for this factor. General reasoning—most constant component of arithmetic-reasoning tests—shows up in many places. The ability is apparently a kind of "all-purpose" or "trouble-shooting" trait, which is called into the picture when more immediate comprehension is lacking.

Conclusions

The principal contribution of this analysis consists in the better definition of mechanical tests. Important also is the additional information concerning factors important in these tests in common with others.

Taking a quick survey of the results, we see that there *is* an important single factor common to printed mechanical tests and also unique in them. It seems clearly to represent a variance in previous mechanical experiences as reflected most clearly in information tests. A strong secondary factor in pictorial mechanical-comprehension tests is visualization. Other factors with substantial loadings are spatial relations and perceptual speed. Deviations from the general pattern are the tests Mechanical Movements (strongest in visualization), Physics (strongest in the verbal factor), Pattern Assembly (strongest in length estimation), and Pattern Comprehension (strongest in general reasoning, though it probably need not be).

All the factors in the list for the mechanical-battery analysis are valid for pilot selection except for the verbal and general-reasoning factors. When any test combines valid factors, its scores are bound to yield exceptionally good predictions. If one desired the most univocal representative of mechanical experience, however, he would choose one of two tests: Mechanical Information and Tool Function. The other valid factors in mechanical tests can be better assessed by means of nonmechanical tests. The unique contribution of tests in this area, therefore, is the mechanical-experience factor. A factor discovered in the analysis of the mechanical battery is that of length estimation, which will be met again in another chapter.

BIBLIOGRAPHY

(1) Bennett, G. K. and Cruikshank, R. M. *A Summary of Manual and Mechanical Ability Tests* (Preliminary Form), New York: The Psychological Corporation, 1942.

(2) Cox, J. W. *Mechanical Aptitude: Its Existence, Nature, and Measurement,* London, England: Machuen, 1928.

(3) Harrell, T. W., A Factor Analysis of Mechanical Ability Tests, *Psychometrika,* 1940, 5, 17–33.

(4) Stenquist, J. L. *Measurement of Mechanical Ability,* Teachers College Contributions to Education, 1923, No. 130.

(5) Thurstone, L. L. *Primary Mental Abilities,* Psychometric Monograph No. 1, Chicago: University of Chicago Press, 1938.

CHAPTER FOURTEEN

Information Tests[1]

RATIONALE OF INFORMATION TESTS

Why Information Tests Are Important

Information tests are not by any means a new type of test. Their use as direct measures of achievement is a long-standing tradition. Only in recent years, however, has their value for indirect measurement of human traits been demonstrated. It is becoming recognized more and more that what a person knows or does not know can be used to reveal a number of things concerning his personal background. Since he is to a large extent a product of his personal experience, and since what he is bodes good or ill concerning his future status in one respect or another, knowledge scores promise to have predictive value. It was therefore decided to exploit information tests as potential predictors of the success of aviation students.

Knowledge as an indication of motivation.—Job descriptions of air-crew positions and studies of students in training indicated that interest or high motivation is important for success. It was hypothesized that the possession of information about a certain area usually accompanies interest in that area, and so information tests were constructed as measures of interest. The quantitative use of self-rating measures of interest, such as are given by the Training Interests Blank, CE501E, in which the student rates his degree of interest in each type of air-crew training, was considered undesirable because of the strong subjective element. On the other hand, it was thought that some graduated evaluation of a candidate's strength of interest should enter into the composite score used in qualifying or disqualifying him for training. Information tests were accordingly intended to provide more objective and reliable measures of interest.

Knowledge as an indication of skills.—An aviation student who has psychomotor aptitudes and skills that are useful in flying will be more likely to succeed in training than a student who lacks such abilities. Information tests were therefore constructed to measure the extent to which candidates have had experiences that develop motor skills pertinent to the air-crew jobs. Trade tests first developed by the United States Army and subsequently by the United States Employment Service use infor-

[1] Written by Staff/Sgt. Benjamin Fruchter.

703320—47—23

mation tests to measure levels of worker skill in each of numerous vocations. From a study of biographical background data it had been found that participation in certain sports and hobbies was prognostic of air-crew success. The measurement of these experiences more objectively than in terms of personal statements called for something like information tests.

Knowledge that may transfer to flying.—A few information tests were designed purely as measures of knowledge. The aviation student who possesses information directly pertinent to a specific job or who has relevant background information so that he can more easily acquire such knowledge will be at an advantage in air-crew training.

Content of Tests to Be Considered

To recapitulate, this chapter deals with tests that use informational items to measure interest, previously acquired skills, experiences pertinent to air-crew jobs, and specific knowledge. These measures of interest or of pertinent experience or knowledge fall into two categories: (1) sports-and-hobbies tests and (2) general-information tests. Tests of mechanical information are treated in the chapter on mechanical tests (ch. 13). Another special group of information tests is described in connection with the assessment of masculine vs. feminine attitudes in chapter 25.

SPORTS AND HOBBIES PARTICIPATION TESTS

Earlier Use in Aviation

During the first World War, the British gave weight to participation in various sports and hobbies in their selective process for military personnel. The United States Army did exploratory work in this area in 1917-1918. In this war, results for the Navy Biographical Inventory suggested a satisfactory predictive value of biographical sports-and-hobbies questions.

Plan of Research

A tentative list of areas to be included in the sports-and-hobbies tests was constructed, utilizing four sources of information:

(1) The Navy blank, a study of which had revealed that certain avocational activities were significant in pilot selection.

(2) Job analyses of the duties of bombardiers, navigators, and pilots, which indicated that participation in certain sports and hobbies should have differential predictive value.

(3) Hypotheses based upon a priori psychological considerations as to which sports and hobbies should be significant.

(4) Information yielded by the results of the administration to 158

aviation students of the Sports and Hobbies Check List, CE506X, of extent of participation in 50 sports and hobbies.[8]

Method of development.—From all sources, information concerning the sports and hobbies to be included was synthesized. Activities were included only if there was either evidence or a sound rationale for their significance and if they were of the kind normally participated in by many air-crew candidates.

All questions developed for sports-and-hobbies tests were of the five-alternative information type. The first four alternatives always included the correct answer and distractors. The fifth answer was always "Don't know." This was included to give the examinee an acceptable way of expressing ignorance of the correct answer and, consequently, avoiding to some extent the chance element involved in guessing.

Sports and Hobbies Participation Test, CE506D and CE506E[9]

These are two equivalent forms of the test. Each form includes 100 items matched with those of the other form for statistical similarity and content.

These questions, with many more, had been given in a preliminary form, CE506C, to examinees who entered pilot classes 44F and 44G for item analysis and item validation, and were selected for inclusion in these equivalent forms on the basis of the following considerations:

(1) The tetrachoric correlation between passing or failing the item and being in the highest 27 percent or lowest 27 percent of the total-score (for a given activity) distribution. The minimum acceptable value of r was +0.60. (There were a few exceptions to this general rule.)

(2) The percentage of responses to the most popular misleads by the upper and lower groups. If not over 20 percent of those in the upper 27 percent of the total-score distribution selected an incorrect answer, the item was included.

Table 14.1 gives the number of items for each activity included in both CE506D and CE506E.

TABLE 14.1.—*Distribution of questions included in each form of the Sports and Hobbies Participation Test (CE506D and CE506E)*

Group 1 (5 items each):	Group 2 (4 items each):
Automobile Driving.	Basketball.
Flying.	Tennis.
Model Plane Building.	Firearms.
Photography.	Hunting.
Radio.	Motorcycling.
Jazz.	

[8] The 50 activities are football, swimming, basketball, baseball, tennis, golf, bowling, fishing, sailing, flying, horseback riding, cards, hunting, model plane building, playing musical instruments, motorcycling, skeet shooting, badminton, dancing, sketching and p[illegible]nting, motion picture (printing and developing), collecting classical music, boxing, ping pong, [illegible]boat racing, skiing, still picture photography (printing and developing), handball, chess, camp[illegible]g, collecting popular records, dramatics, reading, biking, track, writing, marksmanship, pool, radio, ice hockey, leather working, collecting stamps and coins, field events, checkers, motorboat racing, softball, wrestling, fencing, volleyball, debate, a[illegible] metal work.

[9] Developed at Psychological Research Unit No. 1. Chief contributors: T/Sgt. Robert R. Blake and Capt. Donald E. Super.

Group 3 (3 items each):
- Bowling.
- Football.
- Golf.
- Pool.
- Field Events.
- Literature.
- Poker.
- Skiing.
- Sailing.
- Stamp Collecting.
- Diving.

Group 4 (2 items each):
- Chess.
- Riding.
- Woodworking.
- Bridge.
- Dramatics.
- Music appreciation.
- Wrestling.
- Boxing.

Group 5 (1 item):
- Piano.

Two representative items are:

To "draw," a pool player hits the cue ball:

A. At the right.
B. At the left.
C. High.
D. Low.
E. Don't know.

To keep his opponent from getting a combination reverse headlock and arm bar, a wrestler should keep his:

A. Elbows away from his side.
B. Elbows against his side.
C. Arms in front of him.
D. Head down.
E. Don't know.

Keying.—Empirical pilot and navigator keys were constructed on the basis of item validations. Two pilot keys were constructed for each form of the test by separating the odd testing-number papers from the even testing-number papers, thereby obtaining two samples for each form of the test on which to do item validation. For form CE506D the odds group was composed of 422 graduates and 169 eliminees from primary training, and the evens group, of 366 graduates and 170 eliminees. For form CE505E there were 288 graduates and 143 eliminees in the odds group, and 311 graduates and 146 eliminees in the evens group. Examinees were in classes 43I, 43J, and 43K, and they had been tested at Psychological Research Unit No. 1. Responses whose tetrachoric correlation with primary pass-fail data is greater than ±0.12 and whose level of difficulty falls between 90 percent and 10 percent are weighted. In scoring, the odds keys were used to score the papers of the evens samples, and vice versa. The same procedure was followed in obtaining navigator keys, except that only one form of the test (CE506D) was used. In the odds group, there were 302 graduates and 30 eliminees; in the evens group, 300 graduates and 30 eliminees. Examinees were in class 44-9 and had been tested at Psychological Research Unit No. 1. The key, however, was developed at Psychological Research Unit No. 2. Responses whose correlation (phi coefficient) with graduation-elimination from navigator training is greater than ±0.11 and whose level of difficulty is between 85 percent and 15 percent are weighted. Here, also, the odds key was used to score the papers of the evens sample, and vice versa.

Validity.—The keys were tested in the manner described above, the scoring formula for the pilot keys being R−W/3, and for the navigator key, both R − W/3 and R − W. In these formulas, R refers to responses of positive validity and W to responses of negative validity. The formula R−W/3 is, of course, designed to correct for chance success; the formula R − W assays the preponderance of responses of positive validity over those of negative validity. The resultant biserial validities for each pilot key with the elementary pass-fail criterion are given in table 14.2. The biserial validities for each navigator key with graduation-elimination from navigator training is given in table 14.3.

TABLE 14.2.—*Validity of scoring keys of Sports and Hobbies Participation Test, CE506D and 506E, for graduation-elimination from primary training*

Form	Sample	Key	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
CE506D	Odds ..	Evens[2] .	591	0.71	10.16	6.38	6.37	0.36	0.38
CE506D	Evens ..	Odds[3] .	536	.68	4.78	1.30	6.73	.32	.35
CE506E	Odds ..	Evens[4] .	432	.67	5.11	1.29	7.53	.31	.34
CE506E	Evens ..	Odds[5] ..	456	.68	15.06	11.53	7.03	.30	.32

[1] Corrected to an unrestricted stanine standard deviation of 2.00.
[2] The key contains 40 responses of positive validity and 31 of negative validity.
[3] The key contains 33 responses of positive validity and 38 of negative validity.
[4] The key contains 38 responses of positive validity and 45 of negative validity.
[5] The key contains 45 responses of positive validity and 24 of negative validity.

TABLE 14.3.—*Validity of scoring keys of Sports and Hobbies Participation Test, CE506E, for navigator samples*

Sample	Key[1]	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[4]
Odds	Evens[2]	3.93	0.75	4.21	0.39	0.39
Evens	Odds[2]	6.29	3.19	4.79	.33	.34
Odds	Evens[3]	10.08	7.86	3.08	.37	.36
Evens	Odds[3]	12.88	10.48	3.50	.35	.36

[1] In the odds key there are 37 responses of positive validity and 29 of negative validity. In the evens key, the corresponding figures are 31 and 26.
[2] Score is R−W.
[3] Score is R−W/3.
[4] Assuming an unrestricted stanine standard deviation of 2.00.

Validity of specific avocations.—Scores for the various activities in form CE506C were validated separately for pilots in classes 44F and 44G, who had been tested at Psychological Research Unit No. 1. The score for the purposes of this validation is the number correct. The validities are shown in table 14.4.

TABLE 14.4.—*Validation data for the sub-tests of Sports and Hobbies Participation Tests, CE506C, for graduation-elimination of pilots from primary training*

Sub-test	N_t	p_g	M_g	M_e	SD_t	r_{bis}
Auto driving	371	0.62	14.87	12.44	3.86	0.39
Basketball	314	.70	10.47	10.11	5.46	.04
Bowling	270	.67	5.75	6.34	4.57	−.08
Diving	530	.66	4.21	3.80	3.32	.08
Dramatics	164	.58	4.74	5.29	3.09	−.11
Firearms	436	.63	11.08	9.87	4.72	.15
Flying information A	518	.54	9.44	6.90	4.91	.32
Flying information B	374	.50	10.83	7.71	5.73	.34
Football	308	.69	8.91	9.51	5.73	−.06
Golf	386	.69	8.88	8.89	7.97	−.00
Horseback Riding	270	.67	6.83	7.42	3.97	−.09
Hunting	486	.63	9.28	8.02	5.56	.14
Jazz	308	.69	5.89	6.76	5.69	−.09
Model planes	118	.68	3.94	3.58	3.28	.06
Motorcycling	302	.64	3.64	3.20	3.22	.08
Music	118	.68	1.52	2.27	2.35	−.18
Photography	311	.62	2.79	3.59	4.29	−.12
Poker	747	.62	7.55	7.47	4.53	.01
Pool	466	.57	12.80	11.83	5.49	.11
Radio	302	.64	2.70	2.91	3.78	−.03
Reading	287	.63	9.39	10.21	4.71	−.11
Tennis	432	.62	7.73	7.36	4.47	.05
Woodworking	313	.68	7.76	7.21	3.48	.10

Use of the check-list.—A basic question in the development of any participation test is whether the test actually measures participation. Several types of evidence indicate that the sports-and-hobbies tests do measure participation in the activity about which they are concerned. Some information concerning this problem is given in results from a check-list.

(1) *Check-list, CE506X.*—This instrument was administered along with Sports and Hobbies Participation Test, CE506C. The responses to every item were validated against the graduation-elimination criterion. The sample used comprised students of whom 2,052 were later pilot graduates and 1,421 were pilot eliminees from primary training (classes 44F and 44G).

The instructions for the check-list directed the student to indicate the extent of his participation in different activities according to the following scale:

(*a*) Never participated.
(*b*) Rarely participated.
(*c*) Occasionally participated.
(*d*) Frequently participated.
(*e*) Participated to a great extent.

An examination of the results showed that no single alternative yields a high validity. An examination of the differences in correlations from the least degree to the greatest degree of participation, however, reveals a definite trend. The results are given in table 14.5.

(2) *Keying of check-list, CE506X.*—A scoring key for the check-list items was prepared on the basis of the item validation. The key derived from the odd-testing-numbers validation was used to score the papers of the even-numbered sample and vice versa. The odds-sample

Table 14.5.—*Validation of responses to alternatives for stems of the Sports and Hobbies Check List, CE506X, for a sample of 3,473 pilots in primary training (p_e=.59)*

Avocation	r_{tet} for degree of participation: a	b	c	d	e
Automobile driving	(1)	.....	−0.10	0.09	0.10
Basketball	.05	−.01	.00	.02	....
Bowling[2]	−.05	.00	−.03	.09	....
Bridge	.03	.03	−.07	....	....
Chess	.02	.03	.....	....	....
Diving[3]	−.08	.06	.00	....	....
Field events[4]	−.02	.05	−.02	....	....
Firearms[5]	−.10	−.03	.05	.11	....
Fishing	−.11	−.01	.09	.04	....
Football[6]	.08	.02	−.08	.01	....
Golfing	−.05	.00	.02	.09	....
Horseback riding	−.16	.05	.10	.10	....
Hunting[7]	−.17	.....	.15	.06	....
Jazz	.04	.02	−.07	....	....

Avocation	r_{tet} for degree of participation: a	b	c	d	e
Model planes[8]	−0.12	0.07	0.01	......	......
Motorcycling	−.17	.09	.....	......	......
Musical appreciation[9]	.....	.10	−.05	−0.03	......
Photography[10]	−.05	.07	.02	.03	......
Piano playing	.02	.....	.....	......	......
Poker playing	.03	.04	−.01	−.07	......
Radio[11]	.00	.01	.....	......	......
Reading[12]	.....	.15	.06	−.05	−0.20
Sailing	−.04	.01	.03	......	......
Skiing	−.08	.02	.09	......	......
Stamp collecting	.09	−.03	−.10	......	......
Woodcraft[13]	−.06	.05	.04	......	......
Wrestling	.03	−.01	−.01	......	......

[1] A blank indicates a split too extreme to permit computation of a tetrachoric r.
[2] Refers to tenpins only.
[3] Does not refer to swimming.
[4] Jumps, hurdles and pole vaulting only.
[5] Refers to knowledge of firearms.
[6] Actually playing only.
[7] Deer, birds, skeet, and trap shooting.
[8] Flying models only.
[9] Classical music.
[10] Stills and movies.
[11] Construction and use of parts.
[12] Novels, poetry, fiction.
[13] Refers to working with wood only.

key contained 11 weighted items, and the evens contained 11 items. The scoring formula was the number of positively weighted activities checked minus the number of negatively weighted activities checked. The validities are reported in table 14.6.

TABLE 14.6.—*Validity of Sports and Hobbies Check List, CE506X, scoring keys for pilots in primary training*

Group	Key	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
Odds	Evens	1,733	0.60	0.64	−0.27	2.15	0.26	0.26
Evens	Odds	1,740	.58	−2.77	−3.67	2.82	.20	.21

[1] Corrected to an unrestricted stanine standard deviation of 2.00.

(3) *Validation of active vs. sedentary activities.*—The avocations of check-list CE506X were separated into those requiring considerable physical activity and those of a sedentary nature.[4] Students indicated their participation on the five-point scale referred to in the previous section from "never participated" to "participated to a great extent" and were given a cumulative active-sports-participation score and a cumulative sedentary-activities-participation score. The results are shown in table 14.7.

TABLE 14.7.—*Validity data for participation in active sports and participation in sedentary activity as determined by Check List, CE506X, for graduation-elimination from primary training for two samples of pilots*

Group	N_t	p_g	M_g	M_e	SD_t	r_{bis}
Active sports:						
Odds population	1,733	0.60	26.79	25.35	5.88	0.15
Evens population	1,740	.58	26.61	25.74	6.02	.09
Sedentary activities:						
Odds population	1,733	.60	16.52	17.07	3.78	−.09
Evens population	1,740	.58	16.46	17.04	3.6[illegible]	−.10

They indicate that trainees who have engaged in sports and hobbies involving physical activity have a slightly greater chance of graduation from primary training than do those who have engaged in sedentary activities, though the relationship is very low and much less than that for an empirically developed key.

Conclusions.—Knowledge about certain activities (automobile driving, flying, hunting, firearms, diving, model-plane building, and sailing) is related to success in pilot training. These are all outdoor, active avocations requiring the use of the body in various types of coordination. These activities tend to preclude group partcipation, being in each case highly individual for the principal participant. Most of the activities require the use of equipment, the possession or use of which appears

[4] The active sports included basketball, bowling, driving, field events, fishing, football, golfing, horseback riding, hunting, motorcycling, sailing, skiing, and wrestling. The sedentary activities included bridge, chess, jazz, music appreciation, piano, poker, radio, reading, and stamp collecting.

to indicate average or better socio-economic status. Whether this aspect, as such, contributes to validity is unknown.

Knowledge concerning a second group of activities (basketball, horseback riding, poker, pool, skiing, tennis, and track) appears to bear some relationship to the pilot criterion, though the validities of the subtest scores and of specific questions are not great. With the exception of poker and pool, these avocations are also active, outdoor, activities. Poker and pool may be valid because they indicate masculine avocational interests.

A third group of activities (jazz, musical appreciation, piano, photography, radio, and stamp collecting) yields slightly negative validities against the pilot graduation criterion. These activities are largely sedentary and some require thinking, rather than the use of the body, in various types of coordination, and they are usually carried on indoors.

A fourth group of activities (boxing, bowling, football, woodworking, and wrestling) has no relationship with the criterion. This lack of relationship may be due in part to the following difficulties encountered in test construction:

(*a*) The boxing and wrestling questions proved to be generally unsatisfactory. This was in part anticipated during the initial item construction because of regional differences in terminology, rules, and practices for amateur, college, and professional events.

(*b*) The bowling and woodworking items were not satisfactory, probably because of geographical variations in popularity, together with certain terminological variations which made difficult the framing of adequate questions. In bowling, for example, duckpins and tenpins are popular in different sections of the country. Woodworking might well be associated with a range of activities varying from rough carpentering to cabinet making.

(*c*) The football items were unsatisfactory because of the difficulty in developing differentiating questions. Questions about flying that can be answered only by those who have participated in the activity are relatively easy to construct and are meaningful, whereas in football such questions are not easy to write. The unusual popularity of football as a spectator sport makes difficult a clear discrimination between participant and spectator.

Reading, or knowledge of literature, consistently yields a very low negative correlation with the pilot criterion, regardless of the specific method used in measuring participation therein. This negative relationship shows that absorption in this sedentary, abstract avocation is not conducive to success in pilot training.

Generally, questions about how a given end is achieved, when to use a specific approach, or how a thing works, are superior to questions about rules or regulations. It is interesting to note that this differentiation may reflect a difference between participation in and observation of an activity.

Apparently, also, the test may be keyed with equal success for the navigator-training criterion. Unfortunately, no detailed hobby-by-hobby analysis is available for navigators.

A sports and hobbies test, then, appears to be a successful selection instrument. A detailed item analysis, however, is required before such a test can be used for predicting any single criterion variable.

Disposition.—Items from the Sports and Hobbies Participation Test were incorporated in both the AAF Qualifying Examination and the General Information Test, CE505E, of the classification battery.

GENERAL INFORMATION TESTS

This group of tests covers the fields of technical information that would be acquired by those having the interests appropriate to potential success in air-crew positions. These interests are in the areas of aviation information, mechanical information, active sports, navigation, astronomy, gunnery, etc. The items were constructed to indicate participation in, rather than book knowledge of, these subjects.

Technical Vocabulary Test, CE505C [8]

This is the first form of general-information test included in the classification battery, which it entered in July 1942.

Description. (1) *Internal characteristics.*—The test is made up of five-choice vocabulary and information items. Certain items are concerned with planes, plane identification, flying techniques, etc., and they yield a score for pilot information. Others are concerned with astronomy, instruments, maps, etc., and they give a score for navigator information. Still others deal with guns, bomb sights, trajectories, etc., and yield a score for bombardier information. Of the 100 items in the test, 40 are pilot items, 40 are navigator items, and 20 are bombardier items. The following sample items are taken from the pilot, navigator, and bombardier sections of the test, respectively:

The plane with a cannon in its nose is manufactured by:

A. Bell.
B. Boeing.
C. Sikorsky.
D. Douglas.
E. Vultee.

[8] Developed at Psychological Research Unit No. 1. Chief contributors: Maj. R. N. Hobbs and Lt. Col. John S. Thatcher.

Time is usually calculated with reference to:

A. The Naval Observatory in Washington.
B. Zero degrees latitude.
C. Greenwich.
D. The International Date Line.
E. The League of Nations' Observatory in Geneva.

The extent of scatter of bombs around a target is usually expressed in terms of:

A. Angles of divergence.
B. Yards.
C. Probable error.
D. The Error Scatter Pattern.
E. Concentric circles of error.

(2) *Administration.*—Directions are printed on the test booklet, so the test is largely self-administering. The time allowed for the test is 36 minutes, divided into 3 parts. After 12 minutes, all subjects are instructed to go ahead to part II even if they have not finished part I; and after 24 minutes they are instructed to go ahead to part III.

(3) *Scoring.*—There are three scores, one each for the pilot, navigator, and bombardier set of items. The scoring formula for each of the three subtests is R—W/4.

Statistical results (pilot score).—Results will be presented separately for the pilot, bombardier, and navigator sections of the test. The pilot score will be treated first.

(1) *Distribution statistics.*—Distribution data are shown in table 14.8.

TABLE 14.8.—*Distribution data for Technical Vocabulary, CE505C, pilot score, for samples of unclassified aviation students*

N	Date of testing	Psychological research unit No.	M	SD
1,096	December 1942	1	18.8	7.2
2,376	September and October 1942	2	18.5	7.0
1,015	December 1942	2	17.6	7.1
1,143	December 1942	3	17.1	6.9

(2) *Internal consistency.*—The internal consistency of items in the pilot score is indicated by a mean phi of 0.36, with a range from 0.00 to 0.75, and a standard deviation of 0.15, based on the highest 27 percent and the lowest 27 percent of 360 pilots in classes 43K and 44C, who had been tested at Psychological Research Unit No. 3.

(3) *Reliability coefficients.*—Two estimates of reliability, corrected for length, were comparable, as indicated in table 14.9.

TABLE 14.10—*Validity data for Technical Vocabulary—pilot score, CE505C*

Group	Research Unit	Criterion	N_t	p_p	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training Class 43J[2]	1, 2, 3	Graduation-elimination	3,151	0.66	15.68	16.74	7.01	0.17	0.21
Pilots through basic training[2]	1, 2, 3	do	3,046	.57	18.79	17.02	7.02	.16	.20
Pilots through advanced training[2]	1, 2, 3	do	2,978	.54	18.73	17.13	7.02	.14	.19
B-24 pilots in transition training, classes 43J & 43K	1, 2, 3	do	982	.92	19.21	20.16	6.99	−.07	....
B-25 pilots in transition training, classes 43J & 43K	1, 2, 3	do	314	.98	20.4	15.6	7.0	.29	....
B-26 pilots in transition training		do	380	.82	19.89	19.34	6.59	.05	....
Bombardiers, classes 43-8 to 43-11[3]	1, 2, 3	do	1,829	.79	16.00	16.22	6.58	−.02	.00
Bombardiers, classes 43-14 to 43-18[4]	1, 2, 3	do	523	.86	18.4	17.8	7.0	.04	.07
Navigators, classes 43-10 and 43-11	1, 2	do	731	.87	17.67	16.38	7.11	.10	.16
Navigators, classes 43-12 to 43-15	1, 2, 3	do	1,970	.79	17.5	17.4	7.5	.01	.08
Armorers in training	1, 2, 3	Average academic grade	376	....	.....	.....	7.04	[5].10	....
Air mechanics in training Sheppard Field	2	do	232	....	.....	.....	7.32	[5].36	....
Radio operator mechanics in advanced training	1, 2, 3	Graduation-elimination	235	....	.....	.....	7.26	−.25	....
Radio operator mechanics in advanced training	1, 2, 3	Average grades	153	....	.....	.....	....	[5].15	....

[1] Corrected to an unrestricted standard deviation of 2.00.
[2] Same sample, followed through all phases of training.
[3] New aviation cadets, taking 12-week course (no navigation training).
[4] Reclassified pilots, taking 18-week course (some navigation training).
[5] Product-moment correlation.

TABLE 14.9.—*Reliability of Technical Vocabulary, CE505C, pilot score, for samples of pilots*

Type	N	r_{tt}
Odd-even	[1]200	0.82
Correlation of thirds	[2]356	[3].84

[1] Tested at Psychological Research Unit No. 2. Classes 43K and 44C.
[2] Tested at Psycholcgical Research Unit No. 3, November 1942. Classes 43K and 44C.
[3] Median of 3 r's corrected to triple length.

(4) *Difficulty level.*—The difficulty level of pilot items is indicated by a mean proportion of correct responses equal to 0.56, corrected for chance success. The proportions range from 0.17 to 0.97 with a standard deviation of 0.17. These data are based upon results from 400 pilots in classes 43K and 44C, who had been tested at Psychological Research Unit No. 3.

(5) *Factorial composition.*—The pilot score was analyzed with the December 1942 classification battery (N=3,000). Substantial loadings were obtained on the verbal (0.41), pilot-interest (0.34), and mechanical-experience (0.39) factors. Inconsequential loadings were obtained on the perceptual-speed, numerical, spatial-relations, visualization, general-reasoning, and psychomotor-coordination factors. Its communality in the battery was 0.47, which is considerably short of its reliability. For a full description of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Table 14.10 includes typical validities for the pilot score.

(7) *Item validity.*—The validity of responses to pilot score is indicated by a mean phi of 0.03. The range of phis is from −0.05 to 0.16 with a standard deviation of 0.04. These data are based upon responses of 800 graduates and 600 eliminees from elementary pilot training (classes 43K and 44C; originally tested at Psychological Research Unit No. 3), assuming a p_v of 0.50. In general, the items that have the best predictive value deal with plane identification, technical terms related to flying, technical names of planes or plane parts, and experiences related specifically to flying. Items that have low or negative value deal with historical events, the names of scientific inventors, sports records and events, and "book learning" in general.

TABLE 14.11.—*Distribution data for Technical Vocabulary, CE505C, bombardier score, for samples of unclassified aviation students*

N	Date of testing	Psychological research unit No.	M	SD
1,096	December 1942	1	5.0	2.9
2,376	September and October 1942	2	5.8	3.0
1,015	December 1942	2	5.1	3.1
1,143	December 1942	3	4.8	2.7

Statistical results (bombardier score).—Extensive data are available for the bombardier score also.

(1) *Distribution statistics*.—The data are shown in table 14.11.

(2) *Internal consistency*.—The internal consistency of bombardier items is indicated by a mean phi of 0.35 with a range from 0.18 to 0.53 and a standard deviation of 0.09, based on the highest 27 percent and the lowest 27 percent of total scores of 360 pilots in classes 43K and 43C, originally tested at Psychological Research Unit No. 3. A mean phi of approximately 0.25 would be expected by chance in a test of this length (20 items).

(3) *Reliability coefficients*.—Two estimates of reliability, corrected for length, indicate that the bombardier score is very unreliable. See table 14.12.

TABLE 14.12.—*Reliability of Technical Vocabulary, CE503C, bombardier score, for samples of pilots*

Type	N	r_{tt}
Odd-even	[1]200	0.37
Correlation of thirds	[2]356	[3].47

[1] Tested at Psychological Research Unit No. 2. Classes 43K and 44C.
[2] Tested at Psychological Research Unit No. 3. Classes 43K and 44C.
[3] Median of 3 r s corrected to triple length.

(4) *Difficulty level*.—The difficulty level of bombardier items is indicated by a mean proportion of correct responses of 0.34, corrected for chance success. The proportions range from 0.00 to 0.89 with a standard deviation of 0.26. These data are based on 1,400 pilots in classes 43K and 44C, originally tested at Psychological Research Unit No. 3.

(5) *Factorial composition*.—The test was analyzed in the December 1942 classification battery (N=3,000). Its only substantial loadings are on the verbal (0.44) and pilot-interest (0.33) factors in an analysis in which mechanical-experience, perceptual-speed, numerical, spatial-relations, visualization, general-reasoning, and psychomotor-coordination factors are also found. It had a communality of 0.35 in the battery, which exhausted its nonchance variance. For a full description of the factorial composition of this test, see appendix B.

(6) *Validity*.—Validities of the bombardier score for various aircrew and technical specialties are presented in table 14.13. This score proved to be more valid for navigators and pilots than for bombardiers. If the three validities were corrected for attenuation, however, this conclusion might not still hold.

TABLE 14.13.—*Validity data for Technical Vocabulary, CE505C, bombardier score*

Group	Criterion	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training[a]	Graduation-elimination	3,151	0.66	5.25	4.86	2.98	0.08	0.12
Pilots through advanced training[a]	do	2,978	.54	5.22	4.97	3.00	.05	.09
B-17 pilots in transition training[c]	do	1,946	.98	5.0	4.4	2.8	.09	....
B-24 pilots in transition training[c]	do	982	.92	5.00	5.09	2.83	−.02	....
B-25 pilots in transition training[d]	do	314	.98	5.6	5.7	2.7	−.03	....
B-26 pilots in transition training[d]	do	380	.82	5.24	4.70	2.95	.10	....
Bombardiers[c]	do	1,829	.79	5.19	5.07	3.05	.02	....
Navigators[c]	do	731	.87	5.87	5.56	3.13	.05	.04
Navigators[e]	do	1,935	.79	5.7	5.4	2.9	.05	.16
Armorers in training	Average grades	341	....	....	....	2.72	.12[f]	.15
Air mechanics in training[g]	Average grades	232	....	....	....	3.03	.22[f]	....
Radio operator in advanced training[h]	Graduation-elimination	235	....	....	....	2.87	.05	....

[1] Corrected for an unrestricted pilot stanine standard deviation of 2.00.
[a] In class 43J. One sample, followed through advanced training. Tested at Psychological Research Units Nos. 1, 2, and 3.
[b] In classes 43J and 43K. Tested at Psychological Research Units Nos. 1, 2, and 3.
[c] In classes 43-8 to 43-11. Tested at Psychological Research Units Nos. 1, 2, and 3.
[d] In classes 43-10 and 43-11. Tested at Psychological Research Units Nos. 1 and 2.
[e] In classes 43-12 to 43-15. Tested at Psychological Research Units Nos. 1, 2, and 3.
[f] Product-moment correlations.
[g] Tested at Psychological Research Unit No. 2. Entered AAF Technical Training School at Sheppard Field between March and July 1943.
[h] Tested at Psychological Research Units Nos. 1, 2, and 3.

Statistical results (navigator score).—For the navigator score the data are as follows:

(1) *Distribution statistics.*—Distribution data are shown in table 14.14.

TABLE 14.14.—*Distribution data for Technical Vocabulary, CE505C, navigator score, for samples of unclassified aviation students*

N	Date of testing	Psychological research unit No.	M	SD
1,096	December 1942	1	11.2	6.7
2,376	September & October 1942	2	10.3	6.2
1,015	December 1942	2	10.1	6.4
1,143	December 1942	3	10.8	6.4

(2) *Internal consistency.*—The internal consistency of navigator items is indicated by a mean phi of 0.38 with a range from 0.06 to 0.70 and a standard deviation of 0.15. These data are based on the highest 27 percent and the lowest 27 percent in total score of 360 pilots in classes 43K and 44C, originally tested at Psychological Research Unit No. 3.

(3) *Reliability coefficients.*—Two concordant estimates of reliability, corrected for length, appear in table 14.15.

TABLE 14.15.—*Reliability of Technical Vocabulary, CE505C, navigator score, based on samples of pilots*

Type	N	r_{tt}
Odd-even	[1]200	0.79
Correlation of thirds	[2]365	[3].82

[1] Tested at Psychological Research Unit No. 2. Classes 43K and 44C.
[2] Tested at Psychological Research Unit No. 3. Classes 43K and 44C.
[3] Median of 3 r's corrected to triple length.

(4) *Difficulty.*—The difficulty level of navigator items is indicated by a mean proportion of correct responses equal to 0.32, corrected for chance success. The proportions range from 0.00 to 0.89 with a standard deviation of 0.36. These data are based upon results from 360 pilots in classes 43K and 44C, originally tested at Psychological Research Unit No. 3.

(5) *Factorial composition.*—The navigator score of Technical Vocabulary Test, CE505C, was analyzed in three different batteries having a total N of 3,638. Its only significant loading is on the verbal factor (weighted mean of 0.74), and so it appears to be a pure verbal-ability test. Other factors that appeared in these analyses but on which the navigator score had inconsequential loadings are mechanical experience, perceptual speed, numerical, psychomotor coordination, general reasoning, visualization, spatial relations, and length estimation. The weighted mean of the communalities of the navigator score in these three batteries is 0.67. For a full description of the factorial composition of this test, see appendix B.

TABLE 14.16.—*Validity data for Technical Vocabulary, CE505C, navigator score*

Group	Criterion	N_t	p_s	M_s	M_f	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training[2]	Graduation-elimination	3,151	0.66	10.25	9.78	6.22	0.05	0.09
Pilots in primary training[3]	........do	1,520	.75	11.8	9.7	6.5	.19	....
Pilots through advanced training[2]	........do	2,978	.54	10.23	10.01	6.24	.02	.07
B-17 pilots in transition training[4]	........do	1,046	.98	12.3	11.0	6.8	.08	....
B-24 pilots in transition training[4]	........do	982	.92	11.72	11.54	6.81	.01	....
B-25 pilots in transition training[4]	........do	314	.98	12.9	12.1	7.2	.04	....
B-26 pilots in transition training[4]	........do	380	.82	12.27	11.03	7.03	.10	....
Bombardiers[5]	........do	1,829	.79	9.81	9.55	5.76	.03	.05
Navigators[6]	........do	731	.87	16.12	13.36	6.79	.22	.38
Navigators[7]	........do	1,970	.79	15.9	14.3	7.0	.13	.30
Armorers in training	Average grades	376	....	.....	.....	6.11	[8].09	....
Air mechanics in training[9]	........do	232	....	.....	.....	6.31	[8].24	....
Radio operator mechanics in advanced training[10]	Graduation-elimination	235	....	.....	.....	6.83	−.13	....

[1] Corrected to an unrestricted stanine standard deviation of 2.00.
[2] In class 43J. One sample, followed through advanced training. Tested at Psychological Research Units Nos. 1, 2, and 3.
[3] In class 43D. Tested at Psychological Research Unit No. 2.
[4] In classes 43J and 43K. Tested at Psychological Research Units Nos. 1, 2, and 3.
[5] In classes 43-8 to 43-11. Tested at Psychological Research Units Nos. 1, 2, and 3.
[6] In classes 43-10 and 43-11. Tested at Psychological Research Units Nos. 1 and 2.
[7] In classes 43-12 to 43-15. Tested at Psychological Research Units Nos. 1, 2, and 3.
[8] Product-moment correlation.
[9] Tested at Psychological Research Unit No. 2. Entered AAF Technical Training School at Sheppard Field between March and July 1943.
[10] Tested at Psychological Research Units Nos. 1, 2, and 3.

(6) *Validity.*—Because this test was used in the classification battery, extensive validity data are available for it. Table 14.16 gives typical validation results for air-crew and technical specialties. Table 14.17 shows validation data against the criteria of seven navigator grades. For comparison, validation data are given for the pilot and bombardier scores for the same sample and criteria. It is apparent that the test provides a satisfactory navigator-selection score, and since it correlates so slightly with bombardier and pilot criteria, it is also a good classification test.

TABLE 14.17.—*Validity data for pilot, bombardier and navigator scores of Technical Vocabulary CE505C, against the criteria of navigation grades*[1]

Grade	Score	r	r_{corr} [2]
Dead reckoning (ground school)	P	0.09	0.15
	B	.15	.23
	N	.21	.36
Celestial navigation (ground school)	P	−.01	.05
	B	.10	.17
	N	.08	.22
Dead reckoning (flight)	P	.01	.05
	B	.04	.09
	N	.00	.10
Celestial navigation (flight)	P	.03	.08
	B	.02	.09
	N	.04	.17
Meteorology	P	.09	.14
	B	.13	.20
	N	.34	.44
Military	P	−.03	.00
	B	.02	.05
	N	.05	.12
Final composite	P	.04	.11
	B	.11	.20
	N	.15	.31

[1] For a sample of trainees in Hondo classes 43-10 through 43-15. For the bombardier score, a sample of 426 examinees tested at Psychological Research Units Nos. 1 and 2 was used. For the other scores, the sample comprised 463 examinees from Psychological Research Units Nos. 1, 2, and 3.

[2] Assumed unrestricted stanine standard deviation not reported. All r's are product-moment correlations.

(7) *Part-score intercorrelations.*—The three scores intercorrelate as follows: $r_{PN} = 0.20$, $r_{PB} = 0.34$, $r_{BN} = 0.54$. Here it is satisfying to find that the two scores (pilot and navigator) most valid for their own specialties, and at the same time fairly reliable, correlate so little with each other. This makes the two tests excellent classification as well as selection instruments.

Evaluation.—Some of the defects noted in the Technical Vocabulary Test, CE505C, are as follows:

(*a*) The reliability of the bombardier score is so low that it is of questionable value for classification purposes, even when it carries a small percentage of the weight in a composite score.

(*b*) The bombardier score overlaps the navigator score so much (r_{BN} approaches 1.00 when corrected for attenuation) that a separate score has little meaning other than deviation due to sampling errors. Its loading on the aviation-interest factor does indicate, however, that it is not functionally identical with the navigator score.

In view of these considerations, the Technical Vocabulary Test was revised, in the process of which the bombardier score was dropped, and the title was changed to General Information Test, CE505D.

The pilot validities to be expected from the factorial configurations of the three scores are 0.13, 0.11, and 0.19, respectively, for bombardier, navigator, and pilot sections. Averages of obtained pilot validities are 0.12, 0.09, and 0.21, respectively. The closeness of predicted to obtained pilot validities indicate that the full reasons for the latter are known in terms of common factors.

General Information Test, CE505D *

This test, a revision of the Technical Vocabulary Test, CE505C, is designed to measure various types of background information as an indication of interests suitable for training as a pilot or navigator.

Description. (1) *Internal characteristics.*—The revision, besides deleting the bombardier section of the earlier test, extends the pilot section to 60 items, with revision of certain items. The navigator items were left unchanged. The resulting test thus consists of 60 pilot and 40 navigator items. The added pilot items are more concerned with background sports and hobbies than with flying experience. The following item illustrates the new type of material. The other items are of the types illustrated in connection with test CE505C.

On most motorcycles, the throttle is operated by:

A. Pressing a lever on the handle bar.
B. Depressing a foot pedal.
C. Turning one of the handle grips.
D. Depressing a foot pedal for quantity and turning.
E. Don't know.

(2) *Administration.*—The time allowed for the test is 36 minutes, and the 3 parts are separately timed. After 14 minutes all examinees are instructed to go on to part II (items 40-78), even if they have not finished part I (items 1-39), and after 28 minutes they are instructed to go on to part III.

(3) *Scoring.*—The test is scored both for pilot items and navigator items. The score is the number of right responses. For certain items which had been found to bear a negative relationship to pilot success, the incorrect responses in terms of truth or fact are all keyed as rights.

Statistical results (pilot score).—Results will be presented for the pilot score only, since the navigator score is identical with that of form CE505C.

* Developed in the Office of the Air Surgeon, Headquarters, AAF. Chief contributors: Tech./Sgt. Robert R. Blake, Capt. Frederick B. Davis, and Capt. Donald E. Super.

(1) *Distribution statistics.*—Using a sample of 3,000 unclassified aviation students (tested at all three Research Units), a mean of 34.7 and a standard deviation of 6.4 were obtained for the pilot score.

(2) *Reliability coefficient.*—A corrected reliability coefficient of 0.87 was obtained by the odd-even method on a sample of 1,500 unclassified aviation students. This is a satisfying improvement over the reliability for CE505C pilot score (r_{II}=approximately 0.80).

(3) *Factorial composition.*—General Information Test, CE505D, was analyzed in the July 1943 classification battery. The pilot score has loadings of 0.38 on the pilot interest factor, 0.35 on the verbal factor, and 0.30 on the mechanical-experience factor. It has negligible loadings on the perceptual-speed, numerical, spatial-relations, psychomotor-coordination, and general-reasoning factors. This represents a slight improvement in the intended directions—less verbal variance and more pilot-interest variance. Its communality with the other tests in the battery is 0.43. For a full description of the factorial composition of this test, see appendix B.

(4) *Validity.*—Table 14.18 gives the validity of the pilot score for pilot training.

TABLE 14.18.—*Validity data for General Information, CE505D (pilot score), graduation-elimination criterion*

Group	N_t	p_s	M_s	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training[2]	4,779	0.88	35.37	32.67	6.30	0.23	0.25
Pilots through advanced training	7,264	.78	.....	.....	....	.20	.24

[1] Corrected to an unrestricted stanine standard deviation of 2.00.
[2] In class 44E. Tested at Psychological Research Units Nos. 1, 2, and 3.

Evaluation.—The revisions made in the pilot section of this test had the effect of increasing its validity roughly from 0.21 to 0.24. Its most important loadings are on the pilot-interest and mechanical-experience factors. The loading in the pilot-interest factor indicates the value of the test as a measure of interest and motivation. The pilot validity to be expected from the known factors is 0.20. Since the expected and obtained pilot validities are not far apart, we may conclude that all valid factors in the test probably are accounted for. The presence of the verbal variance, even in reduced amount, is of some concern, since this has slight negative validity for the pilot.

The navigator section of the test is the same as in test CE505C, and it continued to yield good results. It was decided to drop the navigator section of the test in the next revision (CE505E), however, for the following reasons:

(*a*) The only significant variance in the navigator score is due to the verbal factor, which is adequately measured by other tests in the battery, e. g., Reading Comprehension Test, CI614.

(*b*) Mathematics Test, CI702E, correlates slightly higher with expressed preference for navigation training (r=0.59) than does the navigator score of General Information, CE505D ($r=0.55$). This indicates that if it is navigation interest that is to be measured by the information test (factorial results fail to exhibit such a factor), then the mathematics test measures it at least as well, and possibly better.

General Information Test, CE505E [1]

When the tests to be included in the November 1943 Classification Battery were selected, it was decided to construct a new General Information Test to replace form CE505D. Accumulated data indicated that items of the following types should be included:

(1) Aviation-information items of the kinds used in General Information Test, CE505D.

(2) Flying-information items developed and validated at Psychological Research Unit No. 1.

(3) Driving-information items of the type developed and validated at Psychological Research Units No. 1 and No. 3.

(4) Mechanical-information items developed and validated at Psychological Research Unit No. 3 (the Mechanical Information Test, CI905A, having been removed from the classification battery).

(5) Technical-vocabulary items developed and validated at Headquarters, Army Air Forces, Office of The Air Surgeon.

(6) Sports-and-hobbies items developed and validated at Psychological Research Unit No. 1 and at Headquarters, Army Air Forces, Office of The Air Surgeon.

Selection of items.—When the items for the final form of Test CE505E were selected, a number of considerations in addition to the anticipated relative sizes of the beta weights were taken into account. Among them were the sizes of the standard deviations of groups of the five kinds of items, the general specifications for the test, and the amount of reliable validity data for the five kinds of items.

The selection of individual items was done, so far as possible, on the basis of individual item-validity coefficients and difficulty indices. Items having the highest validity coefficients and the most appropriate difficulty indices were used. A valid item was not chosen, however, unless the median validity coefficient of all the items of its type was positive. One objective was to use as many valid items as possible at the level of ability represented by the cut-off point for pilot selection by the classification battery. Another aim was to minimize the intercorrelations of the individual items by utilizing items covering a wide range of topics.

[1] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Frederick B. Davis and Capt. Lloyd G. Humphreys.

Description. (1) *Internal characteristics.*—The test is divided into three parts. Part I contains 25 aviation-interest items and has an administration time of 10 minutes; part II contains 32 flying-information items and has an administration time of 12 minutes; and part III contains 43 items of mechanical information, driving information, and sports and hobbies participation, and requires 14 minutes for administration.

Five sample items are shown below, exemplifying the areas of aviation interest, sports and hobbies, mechanical information, driving information, and flying information, in order.

Which one of the following is most commonly used to train pilots on the ground?

A. The Waco Trainer.
B. The Ryan Trainer.
C. The Fairchild Trainer.
D. The White Trainer.
E. The Link Trainer.

The strongest type of construction in skis is called:

A. Concave top.
B. Flat top.
C. Ridge top.
D. Roof top.
E. Don't know.

With pressure on the starter switch, the starting motor runs smoothly, but no contact is made between the starting motor and the engine. The most probable cause of the trouble is that:

A. The armature of the starting motor is loose.
B. The brushes in the starting motor are not making contact with the commutator.
C. The Bendix spring is broken.
D. A fuse is blown.
E. The ignition coil is not functioning properly.

If you were driving along at 50 miles per hour and the right front tire blew out, it would be best to tighten your hold on the steering wheel and:

64-A Step lightly on the brake pedal.
64-B Step hard on the brake pedal.
64-C Turn the wheels slightly to the right.
64-D Disengage the clutch and let the car coast to a standstill.
64-E Turn off the ignition and let the car roll to a standstill in gear.

Using too much bottom rudder in a steep turn will cause the plane to:

A. Slip.
B. Stall.
C. Gain altitude rapidly.
D. Perform a spiral.
E. Don't know.

(2) *Scoring.*—The scoring formula is R−W/4. The test is scored only for pilots.

Statistical results. (1) *Distribution statistics.*—Distribution statistics were obtained for both pre-CTD (College Training Detachment) and post-CTD groups, and are presented in table 14.19. The students while in college detachments received as much as 10 hours of flying experience. This would account for the difference in mean score, at least in part.

TABLE 14.19.—*Distribution statistics for unclassified aviation students on General Information Test, CE505E*

Group	N	M	SD
Post-college[1]	1,500	45.6	12.0
Pre-college[2]	1,920	32.3	13.3

[1] Tested at Psychological Research Units Nos. 1, 2, and 3 with the November 1943 Battery.
[2] Tested at Medical and Psychological Examining Units Nos. 4 to 10 inclusive.

(2) *Internal consistency.*—Since the test is composed of five types of items, a single item-analysis based on total score would be meaningless. Five separate item-analyses, therefore, were made, correlating items with the total score of the group to which the item belongs. The item statistic used was Flanagan's *r* (1). The data are presented in table 14.20.

TABLE 14.20.—*Data on internal consistency for types of items of General Information Test, CE505E based on a sample of 740 classified pilots*[1]

Type of items	Number of items in criterion	M_r	SD_r	Range of *r*
Aviation information	38	0.42	0.16	0.00–0.74
Sports and hobbies	14	.44	.07	.30–.59
Mechanical information	16	.58	.13	.30–.79
Driving information	12	.44	.10	.32–.68
Flying information	19	.43	.09	.21–.64

[1] Tested at Psychological Research Unit No. 3 in November 1943.

The five parts are seen to be quite homogeneous internally. The small number of items in each part enhances the apparent homogeneity, however, and it is difficult to tell how much to attribute to the spurious part-whole correlation.

The five part scores were intercorrelated, and the correlation coefficients were corrected for restriction of range resulting from selection on the pilot stanine. These data are given in table 14.21. There is therefore considerable heterogeneity as between types of items.

TABLE 14.21.—*Intercorrelations of five part-scores of General Information Test, CE505E, corrected for restriction of range*[1] *for a sample of 740 classified pilots*

	1	2	3	4	5
1. Aviation information	....	0.16	0.42	0.24	0.49
2. Sports and hobbies	0.16	....	.32	.26	.18
3. Mechanical information	.42	.32	....	.47	.30
4. Driving information	.24	.26	.47	....	.18
5. Flying information	.49	.18	.30	.18	....

[1] Assuming an unrestricted standard deviation of 2.00 for the pilot stanine.

(3) *Reliability coefficient.*—Two reliability estimates are given in table 14.22.

TABLE 14.22.—*Reliability estimates for General Information Test, CE505E, based upon samples of unclassified aviation students*

Type	N	r'_{II}	r_{II}
Odd-even	[1]1,000	0.77	0.87
Matched halves	[2]500	.75	.86

[1] Tested at Medical and Psychological Examining Unit No. 7.
[2] Tested at Medical and Psychological Examining Unit No. 10.

(4) *Difficulty.*—The difficulty level of items in the test is indicated by a mean proportion of correct responses of 0.48, corrected for chance success. The proportions range from 0.02 to 0.99, with a standard deviation of 0.25. These data are based upon results for 450 unclassified aviation students (pre-college), tested at Psychological Research Unit No. 3 in October 1943.

(5) *Factorial composition.*—The highest loadings for this test are on the mechanical-experience (0.53), verbal (0.43), and perceptual-speed (0.29) factors, in a battery in which spatial-relations, psychomotor-coordination, numerical, mathematics-background, social-science-background, and kinesthetic-motor factors are also found.

No pilot-interest factor was isolated in this analysis. The communality in this battery (November 1943) is 0.65. The matrix of which this test was a part presented many difficulties, and the factorial solution is not entirely satisfactory. These results, therefore, must be taken with some reservations. For a full description of the factorial composition of this test, see appendix B.

(6) *Validity.*—Validity data are available both for total score and part scores. Table 14.23 gives validity for total score for pilots, WASPs, air mechanics, and armament trainees, and table 14.24 gives validities of the part scores and the total score for pilots.

TABLE 14.23.—*Validity data for General Information Test, CE505E*

Group	Criterion	N_t	p_s	M_s	M_e	SD_t	r_{bis}	$_cr_{bis}$
Pilots in primary training[a]	Graduation-elimination	1,676	0.89	50.06	45.35	11.43	0.21	[b]0.35
Pilots in primary training[c]	do	3,146	.84	50.08	46.49	11.40	.17	[b].30
Pilots in primary training[d]	do	504	.92	51.18	46.51	11.76	.20	[e].29
Pilots in primary training[f]	do	406	.93	52.48	52.16	12.01	.19	[e].26
Air Mechanics	Final weighted-average grade	428	....	.....	.....	10.68	[g].23	.....
Armament trainees[h]	Average grades	269	....	.....	.....	12.66	[g].18	.....
WASPs through advanced training[i]	Graduation-elimination	194	.70	.....	.....	.....	.36	.....

[a] In class 44J. Tested at Psychological Research Unit No. 3.
[b] Assuming an unrestricted stanine standard deviation of 2.00.
[c] In class 44I. Tested at Psychological Research Units Nos. 1, 2, and 3.
[d] In class 44H. Tested at Psychological Research Unit No. 2.
[e] Assuming an unrestricted stanine standard deviation of 1.85.
[f] In class 44F. Tested at Psychological Research Unit No. 3.
[g] Product-moment correlation.
[h] In Lowry Field armament classes 34–44A and 35–44A.
[i] In classes 44–W–6 and 44–W–7.

TABLE 14.24.—*Validity data for the five part-scores*[1] *of General Information Test, CE505E, for elementary pilot training*

[N_t=1,076, p_s=.89]

Score	No. of items	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]
1. Aviation information ...	35	20.26	18.61	5.73	0.15	0.25
2. Sports and hobbies	13	3.49	3.26	2.18	.06	.13
3. Mechanical information	20	9.58	8.76	4.46	.10	.21
4. Driving information ...	12	4.55	3.95	2.26	.14	.21
5. Flying information	20	12.74	11.92	2.94	.15	.23
6. Total score	100	50.62	46.50	11.92	.18	.30

[1] These part-scores were obtained after eleven of the items had been reclassified from one part to another on the basis of correlating each item with the five part scores.
[2] Corrected to an unrestricted stanine standard deviation of 1.83.
[3] Tested at Psychological Research Unit No. 3. Class not reported.

Evaluation.—The revisions incorporated in this form raised its validity for pilots to approximately 0.32. It is clear that General Information Test, CE505E, is a highly valuable test in the classification battery. Its validity for pilot training is exceeded by few tests. Since it is a complex, tailor-made test, however, its usefulness is restricted largely to the purpose for which it was constructed.

Its pilot validity is not by any means fully accounted for by its loadings in the mechanical and perceptual factors, though much of it must be so allocated. The pilot-interest factor, which appeared in both forerunners of this test, must have been increased in weight, perhaps to as much as 0.50, judging by the facts. Its present mechanical variance (28 percent, if correctly estimated) is much too high, and that variance is abundantly covered by other tests in the classification battery.

General Information Test, CE505F[4]

This test is a revision of form CE505E.

Informal job analyses of combat flying led to the conclusion that increasing the number of sports-and-hobbies items (especially those negatively keyed), and the number of mechanical-information items would make the test more prognostic of combat success by enhancing variance in an assumed factor of masculinity-femininity.

Description. (1) *Internal characteristics.*—Part I contains 50 items of aviation interest and flying information, and part II contains 60 items of sports-and-hobbies participation. Six mechanical-information items borrowed from the Mechanical-Information test, CI905B, are included in the general information score.

(2) *Administration.*—The examinee is told that if he completes part I before time is up, he may continue with part II. If he has not completed part I when time is up, he is to go on to part II at that time. The time is 20 minutes for each part.

(3) *Scoring.*—The score is simply the number of positively weighted responses. For items where the correct response has negative validity

[4] Developed at Psychological Research Unit No. 3. Chief contributors: Cpl. Albert A. Canfield, Jr., Lt. David H. Jenkins, and Pvt. James A. Walker.

(music, current events, etc.), any mislead is considered the right answer.

Statistical results.—This test was constructed and administered late in the war (beginning September 1944), and so only distribution statistics are available.

(1) *Distribution statistics.*—Using a sample of 470 unclassified aviation students tested in July 1944 at Psychological Research Unit No. 3, a mean of 69.9 and a standard deviation of 10.6 were obtained.

General Information, CE505FX2, GX2 *

These forms consist of 189 items of flying information. They were administered to unclassified aviation students and will form a pool of pre-validated items for use in future revisions of the flying-information section of General Information, CE505F. The items in Part I of each form have only one correct answer. The items of Part II may have several correct answers. Two sample items are presented:

From Part I:

Coordination exercises require:

A. Spirals.
B. Power-off stalls.
C. Pylon eights.
D. Slips.
E. Skids.

From Part II:

An autogiro will:

A. Bank.
B. Roll.
C. Yaw.
D. Pitch.
E. Spin.

Statistical results. The GX2 form was administered to 3,000 unclassified aviation students. Statistical results were not available at the time of this writing. The following results are available for the FX2 form, for examinees tested in April and May 1944 at Psychological Research Unit No. 3.

(1) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.84, corrected for length, was obtained. This figure is based on a sample of 927 unclassified aviation students.

(2) *Test validity.*—A sample of 928 pilots yielded a biserial correlation of 0.37, corrected for restriction of range, between performance in this test and the graduation-elimination criterion in primary training. The mean score for graduates was 63.86, for eliminees 58.26, and the standard deviation for both combined was 12.05. Of this sample, 79 percent was graduates, and the standard deviation assumed for the unrestricted pilot stanine distribution was 2.00.

* Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Charles H. Derthick and Lt. William M. Wheeler.

General Information, CE505FX3, GX1 [10]

These tests consist of 126 aviation-interest items. The questions concern airplanes, tactics, airplane identification, etc. They were administered to unclassified aviation students to provide a backlog of validated items to be used in future revisions of the aviation-interest section of General Information, CE505F.

Statistical results.—The GX1 form was administered to 3,000 unclassified aviation students. Statistical results were not available at the time this volume was being written. The following results were available for the FX3 form, for pilots in class 44J, originally tested at Psychological Research Unit No. 3.

(1) *Reliability coefficient.*—By the odd-even method, an estimated reliability coefficient of 0.81, corrected for length, was obtained. This figure is based on a sample of 939 classified pilots. The score used was rights only.

(2) *Test validity.*—A sample of 927 pilots yielded a biserial correlation of 0.29, corrected for restriction of range, against the graduation-elimination criterion in primary training. The mean score for graduates was 40.51, for eliminees 37.22, and the standard deviation for both combined was 9.06. Of this sample, 68 percent was graduates, and the standard deviation assumed for the unrestricted pilot stanine distribution was 2.00.

General Information, CE505GX8 [11]

This test consists of 65 items designed to assess knowledge of aviation slang. The slang terms are those that would be used on the flight line or in publications on flying. The test was administered to 3,000 unclassified aviation students and will be used as a backlog of prevalidated items for future revisions of the aviation-interest section of General Information, CE505F. Two typical items are:

"Umbrella men" are:

A. Glider pilots.
B. Autogiro pilots.
C. Transport pilots.
D. Paratroopers.
E. Men who have bailed out.

A "mickey" is:

A. A supercharger.
B. An aerial radar unit.
C. A Sperry turn-bank.
D. A pressurized cabin.
E. A droppable gas tank.

[10] Constructed at Psychological Research Unit No. 3. Chief contributors: Cpl. Albert A. Canfield, Jr., T/Sgt. Sanford J. Mock.

[11] Developed at Psychological Research Unit No. 3. Chief contributors: Cpl. Leland D. Brokaw and Cpl. Robert E. Lambert.

Evaluation of General Information Tests

Validities for the tests in this section indicate that the amount of knowledge and information about flying acquired prior to air-crew training is a valid indicator of the background and interest conducive to success in flying training. Analysis showed general-information tests to be factorially complex. In addition to variance in the mechanical-experience and verbal factors, they contained a pilot-interest factor. This was particularly true of the pilot score which was based on questions of (1) flying information, (2) aviation information, (3) sports and hobbies, and (4) driving information. The pilot-interest component of the test proved to be as valid a contributor, perhaps slightly more so, as a measure of motivation for flying training than direct expressions of strength of interest in flying (see ch. 26). Assuming the validity of the pilot-interest factor to be 0.25 (see table 28.17), and the loading of this factor in the test to be 0.50, the factor's contribution to the test validity would be the product of these two values, or 0.125. The validity of self-ratings of pilot interest are generally less than that.

The navigator score was less useful, because it seemed to be duplicating information obtained from mathematics and verbal-test scores.

The bombardier score proved to be too unreliable to be useful for predictive purposes and seemed to contribute nothing unique.

BIBLIOGRAPHY

(1) Flanagan, J. C. *A Table of the Values of the Product-Moment Coefficient of Correlation in a Normal Bivariate Population Corresponding to Given Proportions of successes, New York,* Privately printed, 1936.

CHAPTER FIFTEEN

Perceptual Tests[1]

PERCEPTUAL REQUIREMENTS OF AIR-CREW DUTIES

Even a superficial acquaintance with the duties of the bombardier, navigator, and pilot tempts one to speculate that perceiving—the apprehension of objects and events as present and going on now (1)—is an activity basic to successful air-crew performance. The specification of the important and statistically independent perceptual abilities, however, which would have made possible an economical and searching investigation of the area, was lacking. In the absence of such systematic knowledge, the direction of perceptual research was determined in large part by both formal and informal job analyses. Of these, perhaps the most important analysis, historically speaking, was that of the faculty-board proceedings in the elimination of 1,000 aviation students from further elementary pilot training. This analysis was the source of the classification of perceptual tests, and thus provided the basic framework for test research and construction in this area.

The reasons for elimination stated in the proceedings were placed, upon analysis, into four categories: Coordination and technique, intelligence and judgment, personality and temperament, and alertness and observation. The latter category was taken to coincide with the area of perception. Statements subsumed under this category were found in 70 percent of the eliminations.

The break-down of this gross category of alertness and observation provided the coding system for perceptual test construction. In the list that follows, the first six coded categories and definitions are taken from the published analysis of the faculty-board proceedings; the last two categories were added later. Each perceptual test-code symbol begins with the letters CP followed by a three-place number in one of the following groups.

100. *Visualization of flight course.*—Ability to "get out of the cockpit" and fly the plane with reference to the horizon and reference points, as shown by the ability to handle ground pattern work, maintain constant altitude, control the direction of the plane, make turns of the desired amount, etc.

[1] Written by Capt. John I. Lacey.

200. *Estimation of speed and distance.*—Ability to make such estimates of speed, distance, and altitude as are required in flying a course, flying in formation, gliding, landing, etc.

300. *Sense of sustentation.*—Ability to sense support or lack of support of the airplane, and thus detect slips, skids, or the approach of a stall.

400. *Division of attention.*—Ability of the pilot to remain alert and observant of things around him while flying and at the same time attend to all the necessary details and carry on all the different activities necessary for precise flying.

500. *Orientation.*—Ability to find one's correct geographic position by the use of any available means, such as familiar reference points that are visible on the ground, identification of the area below as it is represented on charts or maps, etc.

600. *Speed of decision and reaction.*—Ability to think quickly, to make rapid decisions, or to respond with speed and precision when the situation demands.

700. *Auditory discrimination.*

800. *Form perception.*

Data on the relative importance of these categories for bombardiers, navigators, and pilots may be found in chapter 1, where it may be seen that perceptual abilities, in general, stand rather high on the lists of required abilities, both in training and combat.

The critical reader may well question why this list, the result of one early job analysis, was accorded such a prominent place in the program of test construction. It is derived from a series of comments made by psychologically naive board members, which were later ordered into a list with some psychological meaning. Its defects as a contribution to a systematic psychology of perception are obvious. The list did have, however, two great advantages. First, it provided a clue to important perceptual activities, tests of which had some promise of validity; and second, it was broad and permissive, providing a flexible if not rigorous scheme of classification. In a time when test construction could not await the detailed job-analysis findings of psychologically informed investigators, these advantages were sufficient to justify the use of these early findings in outlining a program of test construction.

AN OVER-ALL VIEW OF PERCEPTUAL TEST CONSTRUCTION

Without anticipating the detailed discussion to follow, subarea by subarea and test by test, it seems desirable to provide an over-all view of test-construction activity in the field of perception.

It should be noted, first of all, that no tests were constructed in the important areas of sensory psychology. Visual and auditory capacities were the concern of the medical officer, and while civilian research psychologists made important contributions to the field, military psychologists devoted their attention to nonsensory problems.

The coded categories set forth in the preceding paragraphs served as the basic framework, as has been stated, for perceptual test construction. In the chapters to follow, however, an uneven distribution of effort over these categories is apparent. Systematic considerations and relative promise of high validity account for some of this inequality of effort, of course, but it is due mainly to the fact that test-construction activity, especially in the later phases of the program, was directed primarily by validation returns and by factor-analysis results.

As an example of uneven distribution of effort, it was discovered very early that a simple speed test of the ability to match airplane silhouettes was moderately valid for elementary pilot training. This test, Speed of Identification, quite obviously resembled the Identical Forms test of L. L. Thurstone, a test which was known to be heavily saturated with a perceptual-speed factor. It was assumed, and quickly proved in later factor analyses, that the Speed of Identification test was indeed a quite pure measure of the perceptual-speed factor, and almost no further work was done on the test until the very end of the program when a new test, without face validity, was constructed for comparative purposes.

On the other hand, tests sampling spatial abilities were a focus of interest during the entire period of activity in test construction and research. They were known to be valid, but no very pure measures existed. It eventually became clear that at least two spatial factors were involved. Efforts to define these reference abilities more sharply and to determine their validities were still being made at the end of the war.

The organization of the chapters in this section of the volume does not completely follow the coding system. Visualization tests, for example, do not appear at all, since it seemed that they suited an intellectual classification better. The chapter headings represent a compromise between the desire for classification according to primary reference abilities and the present need for a priori classifications in areas yet unexplored with the tool of factor analysis.

RESPONSIBILITY FOR PERCEPTUAL TEST DEVELOPMENT

In the original division of research responsibility in the Aviation Psychology Program, a perceptual research unit was activated in April 1942, at the headquarters of the AAF Training Command. This unit constructed both motion picture and printed aptitude tests for air crew. In October 1943, the responsibility for these two media was divided. A psychological test film unit was activated to continue research with motion pictures, and the responsibility for printed tests of perceptual abilities was transferred to Psychological Research Unit No. 3. In November

703130—47—25

1944, with the deactivation of Psychological Research Unit No. 3 and the transfer of its personnel to Psychological Research Unit No. 2, the responsibility for all printed air-crew aptitude tests was assigned to the latter unit. Research on perceptual tests was also carried on by those concerned with the construction of the AAF Qualifying Examination.[3]

BIBLIOGRAPHY

(1) Bentley, M., *The New Field of Psychology,* New York, D. Appleton-Century Co., 1934.

[3] For a report of this activity, see Report No. 6 of this series. There will be no descriptive treatment in this volume of the perceptual tests constructed for the Qualifying Examination.

CHAPTER SIXTEEN

Perceptual Speed Tests[1]

RATIONALE FOR PERCEPTUAL SPEED TESTS

General Statement

The study of perception is traditional in the history of experimental psychology. In fact, probably no phase of human behavior has been so extensively examined in experimental laboratories. While the known facts concerning the perception of normal human individuals are numerous, the analysis of perceptual activities into fundamental abilities and the measurement of individual differences in those abilities has not kept pace with laboratory studies. Only in recent years has the factor-analysis approach been brought to bear upon the description of separate and distinct abilities in this area.

It does not lie within the scope of this volume to present a review of the work done before the war. It is desirable, however, owing to the fact that the treatment of perceptual tests in this volume is tied up with factorial considerations, to refer briefly to certain previous factorial investigations.

Early Factor Studies

Perhaps the most objective attempt to isolate and to classify the variables involved in perception can be found in Thurstone's work. From a matrix of intercorrelations of 56 psychological tests, he extracted and named seven "primary" factors, one of which was labeled perceptual (factor P). A study of the tests saturated with factor P "indicated that the perceptual factor might consist in a facility to perceive detail even when it is buried among perceptual distractors. . . . The characteristic that seemed to be common . . . was the readiness to discover and to identify perceptual details" (1). This hypothesis was supported by the results of a subsequent analysis, using 22 tests taken from the original 56, plus 9 new tests prepared especially to help define the perceptual factor (2).

In another study (3), Thurstone analyzed the relationship of 43 individually-administered laboratory tests, each of which was designed to measure some aspect of perception. Several new factors were identified as perceptual in nature. These included: Speed and strength of closure

[1] Written by S/Sgt. Wayne S. Zimmerman.

—the ability to hold figures in mind without losing their identity or shape; geometric illusion—the ability to resist the effect of certain geometric illusions; reversal in perception—the tendency to see rapid alternating effects; and freedom from Gestaltbindung—flexibility in manipulating several more or less irrelevant or conflicting Gestalts.

Perceptual Speed in Aviation Psychology

Analysis of the reports of pilot instructors revealed that in 14 percent of all student failures studied a lack of speed in making decisions and in reacting was mentioned as a cause for elimination. The trait of "speed of decision and reaction" was described as the ability to think quickly, to make rapid decisions, and to respond with speed and precision when the situation demands. Some typical comments by the instructors regarding eliminated students were: "Slow to think and act in the air," "suffers from indecision," "unable to make rapid decision," "choice of fields slow and unsatisfactory," and "slow reaction time." No specific mention was made of speed of perceiving or apprehending a path or a pattern, or of speed of distinguishing meaningful visual detail, although the lack of such abilities may be partially responsible for remarks such as those just cited.

In the report of student failures referred to in the preceding paragraph, lack of judgment was the most frequently listed cause of elimination. To what degree evaluations of judgments as good or poor depend upon quickness of decision is not known. It is reasonable to assume, however, that slowness in perception is a contributing factor.

Tests that were found to measure perceptual speed are described and discussed in this chapter under the subheadings (1) Speed of Apprehending Perceptual Detail, and (2) Clerical Speed.

TESTS OF SPEED IN APPREHENDING PERCEPTUAL DETAIL

Tests described within this group have in common problems that appear to demand the rapid visual perception of detail or the recognition of similarities and differences. A comparison of form and design and an identification of patterns or details that may be buried among perceptual distractors are involved. The individual items are simple; a good score depends almost entirely upon the rapidity with which the examinee can perceive the details.

Speed of Identification, CP610A [1]

This test was designed to measure speed and accuracy of form perception. The speed with which identical airplane silhouettes can be identified by quickly noting differences and similarities of form was believed to be indicative of the prospective air-crew member's ability to

[1] Developed at the Office of The Air Surgeon, Headquarters, Army Air Forces. Chief contributors: Lt. Col. Paul M. Fitts and Staff.

perform rapidly the task of identifying enemy aircraft, taking instrument readings, noting airplane altitudes, recognizing landmarks, and accomplishing other activities demanding speed and precision of perception.

Description.—The test sheet is divided into 12 panels. On the left in each panel are silhouettes of four airplanes. On the right are silhouettes of five airplanes, four of which are identical with the four on the left and one of which is different. The planes at the right are rotated, in a haphazard manner, into different positions than those at the left. For each plane on the left, the examinee must find the matching one on the right. The test is printed directly on an IBM answer sheet. Under each plane in the left of a panel are answer spaces marked A, B, C, D, and E. The five planes on the right are labeled A, B, C, D, and E to correspond. Thus, if plane B on the right is identical with the top plane on the left, that top plane should have space B blackened below it.

(1) *Internal characteristics.*—A practice-test sheet is provided, containing one panel of four recorded, but unscored, practice items. The test contains 12 panels, with a total of 48 recorded and scored items printed on 2 sides of an IBM answer sheet.

(2) *Administration.*—After the directions are read and the signal to begin the test is given, the perforated directions sheet is torn away from

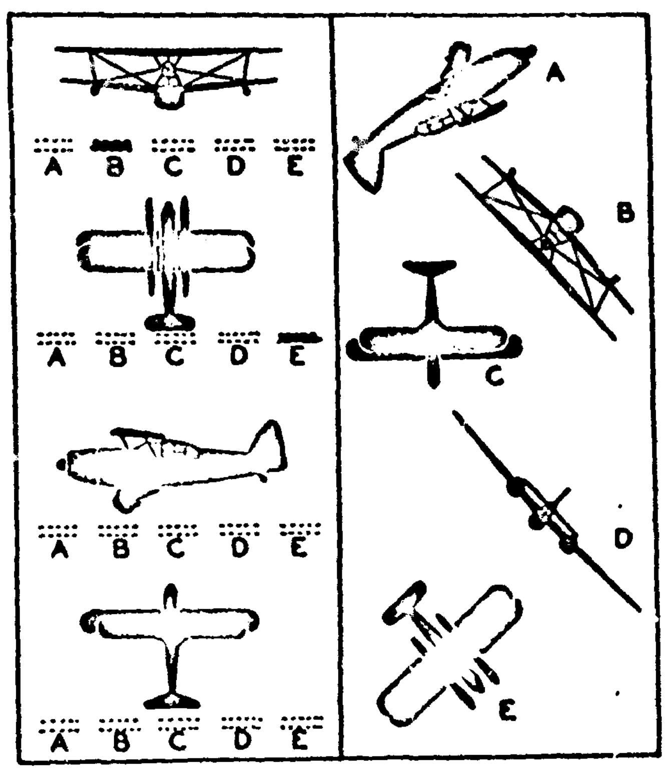


FIGURE 18.1
SAMPLE PROBLEM OF SPEED OF IDENTIFICATION, CP810A

the test sheet by the examinee. The administration time is 5 minutes, while 4 minutes are allowed for completing the test items, making a total testing time of 9 minutes. Practice items are shown in figure 16.1.

(3) *Scoring.*—From 1 December 1942 to 10 July 1943, this test was scored (R − W)/2. Both before and after that period the scoring formula was R — W.

Statistical results.—Having been a classification-battery test, Speed of Identification was extensively analyzed.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 16.1. The distribution curves are negatively skewed and considerably flatter than normal.

TABLE 16.1.—*Distribution constants for Speed of Identification, CP610A*

Group	Psychological research unit No.	N	M	SD	Scoring formula
Unclassified aviation students	1, 2, 3	3,000	33.3	7.3	R−W
Do	2	1,520	25.6	6.9	R—W
Do	1	2,729	31.3	7.5	R−W
	1	1,096	14.9	3.8	(R−W)/2
Classified pilots[1]	1, 2, 3	972	31.6	7.6	R—W
Classified navigators	3	367	32.6	7.5	R−W
Radio operators[2]	1, 2, 3	153	14.7	3.9	(R−W)/2
Armorers[2]	......	120	14.4	3.8	(R−W)/2
West Point cadets[3]	......	888	38.7	7.2	R−W

[1] Classes 43J and 43K.
[2] Previously eliminated from pilot training.
[3] Class of 1946.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.09, a standard deviation of the phi distribution of 0.09, and a range of values from −0.09 to 0.36. The phi values are low, because the test is highly speeded. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 700 unclassified aviation students tested in April 1942 at Psychological Research Unit No. 3.

(3) *Reliability coefficients.*—Several samples yielded the estimates of reliability given in table 16.2.

TABLE 16.2.—*Reliability coefficients for Speed of Identification, CP610A, based upon groups of unclassified aviation students*

N	Type	r'_{tt}	r_{tt}
[1]1,020	Alternate forms[2]	0.61	0.76
[3]144	Test-retest (24 hours approximately)[4]	....	.49
[3]321	Do[5]	....	.63
[6]158	Do[7]	....	[8].23
712	Test-retest (30 days)	....	.57

[1] Tested at Medical and Psychological Examining Unit No. 8. Date of testing not reported; data reported in May 1945.
[2] For experimental purposes the test was divided into separately timed halves.
[3] Tested at Psychological Research Unit No. 1 in June 1942.
[4] First administration, 4 minutes, second administration, 3 minutes.
[5] First and second administrations, 4 minutes.
[6] Tested at Psychological Research Unit No. 1 in August 1942.
[7] First administration, 4 minutes; second administration, 6 minutes.
[8] Extremely low correlation due to many cadets finishing all items in the 6 minute retest.

TABLE 16.3.—*Validity data for Speed of Identification, CP610A, for pilots in training, using the graduation-elimination criterion*

Group	Psychological research unit No.	Class	Scoring formula	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
In primary training	2	42I–43A	R−W	6,110	0.62	24.7	22.4	7.0	0.20	....
Do	2	43D	R−W	1,520	.75	26.3	23.5	6.9	.24	....
Do	2	43E	R−W	1,148	.76	25.7	23.8	6.7	.17	....
Do	1, 2, 3	43J	(R−W)/2	3,151	.66	15.7	14.6	3.65	.18	0.23
Through basic training	1, 2, 3	43J	(R−W)/2	3,046	.57	15.8	14.7	3.64	.18	.23
Through advanced training	1, 2, 3	43J	(R−W)/2	2,978	.54	15.8	14.8	3.65	.18	.23
In primary training	1, 2, 3	44E	R−W	4,779	.88	34.04	32.40	7.12	.12	.15
In B-17 transition training	1, 2, 3	43J–43K	(R−W)/2	1,046	.98	16.5	15.2	3.5	.15	....
In B-24 transition training	1, 2, 3	43J–43K	(R−W)/2	982	.92	15.86	15.91	3.66	−.01	....
In B-25 transition training	1, 2, 3	43J–43K	(R−W)/2	313	.98	16.3	17.4	3.6	−.13	....
In B-26 transition training	1, 2, 3	43J–43K	(R−W)/2	380	.82	16.27	15.80	3.50	.08	....
In P-40 assignment	1, 2, 3	43J–43K	(R−W)/2	2,416	.64	16.3	16.1	3.7	.03	....

[1] Assuming an unrestricted stanine standard deviation of 2.00.

TABLE 16.4.—*Validity data for Speed of Identification, CP610A, for navigators in training, using the graduation-elimination criterion*

Group	Psychological research unit No.	Class	N	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
New students	1	42-10 through 42-17	228	0.79	[2]30.3	27.9	8.2	0.16	....
Reclassified pilots	1	42-10 through 42-17	183	.84	[2]30.7	26.8	8.5	.25	....
New students	2	43-11 through 42-16	163	.71	[2]23.8	21.4	6.8	.21	....
Reclassified pilots	2	42-11 through 42-16	392	.77	[2]24.0	21.5	6.1	.24	....
New students and reclassified pilots	3	[3]	367	.90	[2]32.8	28.3	7.5	.31	....
New students	1, 2	43-10, 43-11	725	.87	[4]14.7	13.9	4.3	.10	0.15
Do	1, 2, 3	43-12 through 43-15	1,969	.79	[4]15.5	14.5	3.8	.16	.20

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Scored R−W.
[3] Not reported. Testing dates: Apr. 1 to Aug. 14, 1942.
[4] Scored (R−W)/2.

(4) *Difficulty.*—Based upon the responses of the above-mentioned sample of 700 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.94, corrected for chance, with a range from 0.10 to 0.99, and a standard deviation of 0.08.

(5) *Factorial composition.*—The most significant loading is in the perceptual-speed (0.54) factor in which the test is almost pure. The communality is 0.63. For a full picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Validation results based on several samples are given in tables 16 3 to 16.7 inclusive.

TABLE 16.5.—*Validity data for Speed of Identification, CP610A, using seven grades in navigation training as criteria, for a sample of 463 navigation trainees in Hondo classes 43-10 through 43-15*

Grade	r[1]	r_{corr}[2]		r[1]	r_{corr}[2]
Dead reckoning (ground school) ..	0.06	0.11	Meteorology	0.09	0.13
Celestial navigation (ground school)	.08	.12	Military	−.08	−.06
Dead reckoning (flight)	.11	.14	Final composite	.09	.15
Celestial navigation (flight)	.08	.12			

[1] Product moment correlations.
[2] Assumed unrestricted stanine standard deviation not reported.

Variations.—The B form of Speed of Identification was to have been a lantern-slide adaptation using the items of the original test. It was suggested that tachitoscopic exposures would result in a purer measure of perceptual spee than the temporally uncontrolled printed administration. This slide adaptation was never fully developed, so data to verify or disprove the hypothesis are not available.

The C form was prepared for a two-fold purpose:[3] (1) To remove the possible influence of aviation interest by constructing items utilizing meaningless symbols rather than airplanes, and (2) to reduce the item difficulty as close to zero as possible in order to provide a pure speed test.

For the purposes of factor analysis, a form of Speed of Identification was constructed with the response alternatives unrotated from their original position.[4] This version is without a code designation.

Evaluation.—Speed of Identification, CP610A, was used in the classification battery from March 1942 to November 1943. It was dropped at that time, because it became known that Spatial Orientation I measured perceptual speed nearly as well and, furthermore, was more valid for pilots, due to additional factor content. Speed of Identification was reinstated in September 1944 because of a recognition of the fact that it was the strongest and purest measure of the perceptual-speed factor thus far constructed. Although its average primary pilot validity of 0.18 is lower than that of either Spatial Orientation I or II, its higher loading on the perceptual-speed factor indicates that practically all of its validity is due

[3] Developed at Psychological Research Unit No. 3. Chief contributor: Pfc. Sidney W. Finkel.
[4] Constructed at the Psychological Section, Headquarters, AAF Training Command, Fort Worth, Tex.

TABLE 16.6.—*Validity data for Speed of Identification, CP610A, for bombardiers in training, using the graduation-elimination criterion*

N	Part	Psychological research unit No.	Class	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
[2]416	Front	1	43–5 through 43–7	0.84	20.6	19.3	4.5	0.16	....
[2]416	Back	1	43–5 through 43–7	.84	6.1	4.2	5.7	.19	....
[2]329	Front	2	43–5 through 43–7	.86	21.5	20.6	3.6	.14	....
[2]329	Back	2	43–5 through 43–7	.86	5.8	3.0	5.9	.25	....
[2]469	Front	3	43–5 through 43–7	.82	21.9	21.7	3.0	.03	....
[2]469	Back	3	43–5 through 43–7	.82	11.3	9.7	7.2	.13	....
[3]1,829	Total	1, 2, 3	43–8 through 43–11	.79	15.3	15.1	3.8	.03	0.06
[4]456	........do	1, 2, 3	43–14 through 43–18	.84	15.9	15.6	3.8	.04	.07
[4]524	........do	1, 2, 3	43–14 through 43–18	.86	16.6	16.9	3.5	–.04	.01

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] New students taking the 12-week course (no navigation training).
[3] New students taking the 18-week course (some navigation training).
[4] Reclassified pilots taking the 18-week course.

TABLE 16.7.—*Validity data for Speed of Identification, CP610A, using various criteria*

Group	Class	Psychological research unit No.	Criterion	N	r
Pilots in advanced training	43J	1	Gunnery: Total percent hits	459	0.08
Do	43J	1	Gunnery: Qualification percent hits	459	.02
Do	43K	1	Gunnery: Total percent hits	403	.06
Do	43K	1	Gunnery: Qualification percent hits	403	.03
Pilots in advanced single-engine training	43J–44A	3	Proficiency rating	562	.14
Pilots in advanced twin-engine training	43J–44A	3	........do	685	.01
Pilots in four-engine transition training	43J–44A	3	........do	685	–.01
Bombardiers in training	42–11 through 42–16	2	Combat circular error	180	–.01
Do	42–11 through 42–16	2	Record circular error	180	–.03
Do	42–11 through 42–16	2	Combat circular error	307	.18
Do	42–11 through 42–16	2	Record circular error	307	.11
Air mechanics in training		1, 2, 3	Average grades	232	.27
Armorers in training		1, 2, 3	........do	376	–.01
Radio operator mechanics in training		1, 2, 3	........do	153	–.09
Radio operator mechanics in training		1, 2, 3	Graduation-elimination	235	.20
Flexible gunners			Academic grades	148	.27
			Ground to ground firing	148	.14
			Air to air firing	148	.10
Do	43–45	1, 2, 3	........do	194	–.01
Do	43–45	1, 2, 3	Final examination	194	.26
Do	43–48	1, 2, 3	Air to air firing	171	.03
Do	43–48	1, 2, 3	Final examination	171	.24

to saturation with this factor. Loadings with other factors (measured better by other tests) account for the slightly higher validities of Spatial Orientation I and II.

In an analysis which included both the rotated and nonrotated forms of Speed of Identification, no significant differences in factor structure between the two forms were revealed (see table 28.15). It might be expected that rotation would increase visualization content, and there is some slight indication that this is so; but apparently the perceptual differences in design of the airplanes are sufficiently gross that the examinee finds it unnecessary to rotate mentally an image of the object in order to make comparisons. Thus, only speed of perception is involved to a significant extent.

Pursuit Test, CP414A [5] (Path Tracing, CP512A)

It was thought that some elements of foresight and planning might be involved in tests like the McQuarrie Path Tracing Test. A modification of this test was therefore prepared for group-test administration to aircrew candidates.

Description.—Items are arranged in blocks of 10. Down the left-hand side of each block are 10 numbers, and down the right-hand side of the block are 10 letters. Each number is connected by an irregularly curved line to a letter on the opposite side of the block. Thus a maze of lines is formed. The examinee's task is to trace visually each line from its beginning to its termination and to mark the appropriate letter opposite the item number on the separate answer sheets.

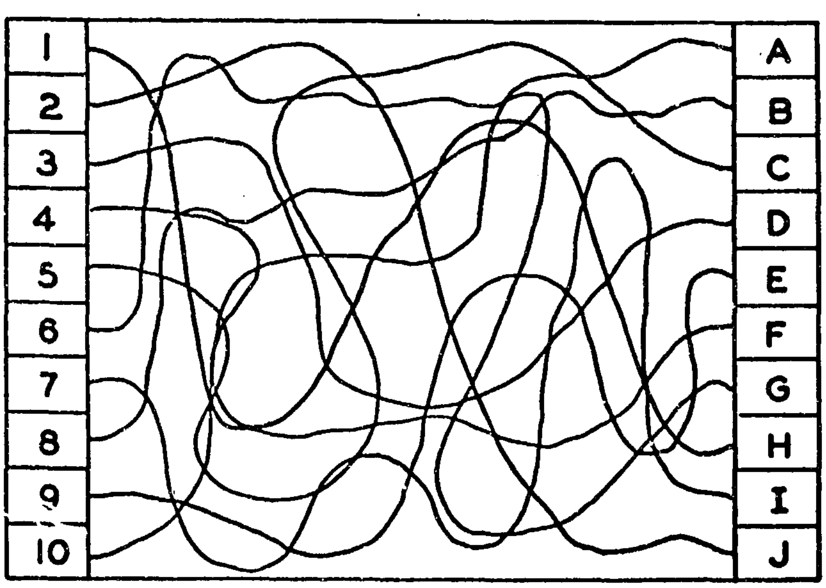


FIGURE 16.2
SAMPLE PROBLEMS OF PURSUIT TEST, CP414A

[5] Developed at Psychological Research Unit No. 3.

(1) *Internal characteristics.*—The Pursuit Test is divided into two parts, separately assembled. It is mimeographed. The Path Tracing test is the printed version of Part I of the Pursuit Test. The directions contain 10 recorded, but unscored, practice items. Each part of the test contains 80 scored items, in 8 blocks of 10 items each.

(2) *Administration.*—Answers are marked directly on two 15-place answer sheets. Five minutes are allowed to complete 80 items. The sample items are shown in figure 16.2.

Following is part of the directions:

In the example, you will find 10 lines which run across the diagram from left to right. The beginning of each line is numbered. Notice that line 1 begins in the upper left corner and ends at the letter H. Opposite item 1 on your answer sheet blacken the space under H.

Similarly, notice that line 2 ends at letter J and that line 3 ends at letter D. For items 2 and 3 on your answer sheet, blacken the space under J and D respectively.

(3) *Scoring.*—The scoring formula is R−W.

Statistical results.—Data are available for this test on small samples. With the exception of the reliability coefficient, all data are for Path Tracing, or for Part I of the Pursuit Test.

(1) *Distribution statistics*—Distribution constants are presented in table 16.8.

TABLE 16.8.—*Distribution constants*[1] *for Pursuit Test, CP414A, and Path Tracing, CP512A*

Form	Group	N	M	SD
CP414A	Classified pilots[2]	199	46.5	6.8
CP512A	Unclassified aviation students[3]	460	47.3	7.8

[1] Scored rights only.
[2] In class 44A. Tested at Psychological Research Unit No. 3.
[3] Tested at Psychological Research Unit No. 3 in March 1943.

(2) *Reliability coefficients.*—By the alternate-forms (part I–part II) method, an estimated reliability coefficient of 0.80, corrected for length, was obtained. This figure is based on a sample of 210 unclassified aviation students tested in April 1942 at Psychological Research Unit No. 3.

(3) *Factorial composition.*—For Path Tracing, CP512A, the most prominent loadings are in the perceptual-speed (0.51), planning (0.27), numerical (0.25) and spatial (0.17) factors. The communality is 0.50, to be compared with the uncorrected reliability of 0.66 for the Pursuit test. For a full pictrue of the factorial composition of this test, see appendix B.

(4) *Test validity.*—Validation results are given in table 16.9.

TABLE 16.9.—*Validity data for Pursuit Test, CP414A, and Path Tracing CP512A, based upon graduation-elimination of pilots in primary training*

Form	Class	Psychological research unit No.	p_e	N_t	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$
CP414A	44A	3	0.86	199	46.63	45.67	6.84	0.08	.....
CP414A	43J	3	.82	123	41.12	41.32	6.86	−.02	.....
CP512A	44A, 44B, 44C	1	.74	640	47.91	46.08	7.56	.14	[1]0.17
CP512A	43K	3	.87	186	54.60	56.38	7.28	−.13	.....

[1] Assumed unrestricted stanine standard deviation not reported.

Evaluation.—The average pilot validity of Pursuit, CP414A, is only 0.09. The factor content of Pursuit is puzzling. In Thurstone's analysis of 56 tests, Pursuit appeared to be highly saturated with factor S (space). In a re-analysis by aviation psychologists (not covered in this report) of 19 of the 56 tests, using the same intercorrelations, the same factor picture was derived for this test. Results of several subsequent analyses based on form CP512A, however, showed most of its common variance on the perceptual-speed factor. One rationale advanced to explain the perceptual content is that speed and accuracy of response is gained by rapidly perceiving the details of a path. The criss-crossing of pathways requires considerable close examination and the seeing of patterns in spite of entangling distractors.

Evaluation of Tests of Speed of Apprehending Detail

The similarity of Speed of Identification and Thurstone's Identical Forms test is substantial evidence for believing that his factor P and the perceptual-speed factor found in AAF tests are one and the same. Why the Pursuit test should appear to be primarily a spatial test in one analysis and a perceptual-speed test in another analysis is not answered. Evidence that it belongs with perceptual speed rather than with spatial tests is strongly supported by three separate analyses, each based on data collected from independent samples. It is possible that this type of test is very sensitive to minor changes of design; i. e., that alterations in drafting the pathways may call for distinct functional shifts in the task of examinees.

Average pilot and navigator validities for Speed of Identification of 0.18, and other evidence, suggest that a conservative estimate for the validity of the perceptual-speed factor would be between 0.15 and 0.20 for both air-crew specialties. These findings bear out the original prediction that measurements of perceptual ability would be a valuable aid to predicting air-crew success.

CLERICAL SPEED TESTS

Tests described within this group are similar in that the tasks involved are clerical. No test specifically designed to measure general clerical ability is included, but the individual problems are similar to or are related to clerical tasks. The problems include reading graphs, dials, meters, or tables, and checking, classifying, or filing numbers. Support for including tests of this nature may be found in the average ratings made by supervisors of combat teams, which show that on a nine-point scale, dial-and-table reading has a mean rating of 6.8 for pilots (see table 1.6). Ratings for combat navigators give dial-and-table reading a mean rating of 6.6 (see table 1.4).

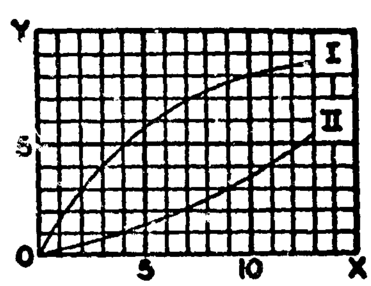
Graph Reading, CP601B

This test and the next four tests described in this chapter are Parts I through V of the Quantitative Perception Tests copyrighted in 1941 by the Cooperative Test Service. These tests were adopted and used for a short time in the classification battery during the first months of testing. When the tests were selected, it was too early in the program for validation data to be available. They were appealing, because the test items presented tasks that air-crew members were known to encounter to some extent in their training and in later operations. They were replaced by apparently better designed tests of similar content or function before validation data were known.

Description.—The answers to all of the test problems are read from a graph on which are drawn two curves. X and Y axes are labeled, and values are indicated on abscissa and ordinate. Each problem gives either the X or Y value for one of the curves, and the examinee is required to determine the other value. Two alternative answers from which the correct one is selected are provided for each item.

(1) *Internal characteristics.*—The directions contain one graph from which the answers to eight items are read. Four of the answers are indicated correctly, and four are recorded by the examinee but are unscored. The test contains 1 graph from which the answers to 52 recorded and scored items are read.

(2) *Administration.*—Each examinee receives the special IBM form on which are presented both directions and items. The directions are printed on one half of one side of an IBM answer sheet. The items are presented on the other half of the page and are arranged so that they appear upside down while the directions are being read. Thus, the page must be turned end for end when the signal to begin the test is given. This format and procedure apply to the other tests in the series. Three minutes are allowed to complete the 52 test items. Sample items are shown in figure 16.3.



CURVE I			CURVE II		
X	Y		X	Y	
2	3	4	7	2	3

FIGURE 16.3
SAMPLE PROBLEMS OF GRAPH READING,
CP601B

Following is part of the directions:

Read the y-values corresponding to the values of x given below for curve I to the nearest whole number; repeat for curve II. The spaces under the correct answers have already been marked in the sample. (See fig. 16.3.)

The correct answers may be obtained in the following manner: (1) Find the point on the x axis (the heavy horizontal line marked OX) where x is equal to 2; (2) follow the heavy vertical line upwards until you find the point that it crosses curve I; (3) from this point follow across horizontally to the y axis (the heavy vertical line marked OY) and read the y value, which is 3 in this case. In the sample, the answer space under 3 has therefore been marked.

(3) *Scoring.*—The first scoring formula was R−3W. It was changed to R — 2W, and finally to R — W.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 16.10.

TABLE 16.10.—*Distribution constants for Graph Reading, CP601B*[1]

Group	N	M	SD
Unclassified aviation students[2]	1,134	18.3	7.4
Unclassified aviation students[3]	226	16.3	6.2
Classified navigators[4]	365	19.2	6.8
Classified bombardiers[4]	557	16.4	7.4

[1] On these samples scoring formulas could not be determined from available data.
[2] Group not identified.
[3] Tested in March and April 1942 at Psychological Research Unit No. 1.
[4] Tested in the period Apr. 1 to Aug. 14, 1942, at Psychological Research Unit No. 3.

(2) *Reliability coefficient.*—By the odd-even method, an estimated reliability coefficient of 0.70, corrected for length, was obtained using the scoring formula R−3W. Owing to the fact that the test is highly speeded, this figure is spuriously high. The estimate is based on a sample of 226 unclassified aviation students tested at Psychological Research Unit No. 1 in March 1942.

(3) *Test validity.*—Validation results based on several samples are given in table 16.11.

Evaluation.—Graph Reading proved to have high validity for navigators and a moderate validity for pilots.

The reliability, although low, is acceptable for inclusion in a battery. The test is comparatively easy to administer and to score.

Meter Reading, CP602B

This is Part II of the Quantitative Perception tests.

Description.—Each item consists of a diagram of a portion of a meter with a needle indicating a reading. The examinee is required to read each dial to the nearest whole number. In nearly all items some of the divisions between numbers are removed and the examinee must estimate needle positions. Two alternative answers are listed for each item.

(1) *Internal characteristics.*—The directions contain five items with the correct answers already marked and five recorded, but unscored, practice items. The test contains 50 recorded and scored items.

TABLE 16.11.—*Validity data for Graph Reading, CP601B*

Group	Criterion	N_t	r_p	M_p	M_e	SD_t	r_{bis}
Pilots in primary training[1]	Graduation-elimination	1,134	.....	.....	.....	.....	0.20
Do[1]	........do	2,505	.....	.....	.....	.....	.16
Do[1]	........do	1,848	0.74	[2]4.76	[2]4.14	[2]1.81	.20
Navigators (new cadets)[3]	........do	228	.79	17.3	12.6	7.3	.37
Navigators[4]	........do	365	.90	20.5	14.7	6.8	.44
Do[5]	........do	163	.71	17.7	1[illegible].4	7.2	.53
Navigators (reclassified pilots)[3]	........do	183	.84	18.9	14.2	7.3	.36
Navigators (reclassified pilots)[6]	........do	392	.77	19.5	14.5	6.5	.46
Bombardiers[4]	Record circular error[7]	557	.....	.....	.....	.....	[8].07
Do[9]	........do[7]	194	.....	.....	.....	.....	[8].06

[1] Groups not identified.
[2] Using scaled scores with a mean of 5.00 and a standard deviation of 2.00.
[3] In classes 42-10 through 42-17. Tested at Psychological Research Unit No. 1.
[4] Tested at Psychological Research Unit No. 3 in the period April 1 to August 14, 1942.
[5] In classes 42-11 through 42-16 at Hondo Army Air Field. Tested at Psychological Research Unit No. 2.
[6] Tested in February, March and April 1942 at Psychological Research Unit No. 2.
[7] Highly unreliable criterion.
[8] Product-moment correlation.
[9] In classes 42-12 through 42-16. Tested at Psychological Research Unit No. 1.

(2) *Administration.*—Each examinee receives a special IBM form on which both directions and items are printed. Three minutes are allowed for completion of the test items. Sample items are shown in figure 16.4.

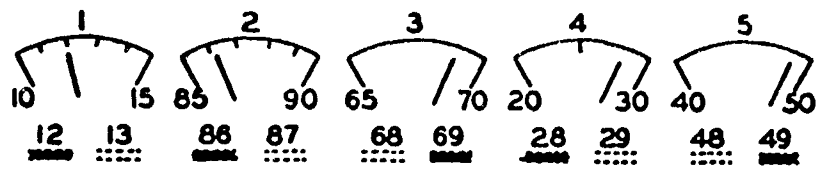


FIGURE 16.4
SAMPLE PROBLEM OF METER READING, CP602B

(3) *Scoring.*—The first formula used was R−3W. It was changed to R — 2W, and finally to R — W.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 16.12.

TABLE 16.12.—*Distribution constants for Meter Reading, CP602B*[1]

Group	N	M	SD
Unclassified aviation students[2]	234	22.2	10.6
Classified bombardiers[3]	557	20.7	10.6
Classified navigators[3]	365	24.6	10.0

[1] On these samples scoring formulas could not be determined from available data.
[2] Tested in March and April 1942 at Psychological Research Unit No. 1.
[3] Tested in the period April 1 to August 14, 1942 at Psychological Research Unit No. 3.

(2) *Reliability coefficient.*—By the odd-even method, an estimated reliability coefficient of 0.73, corrected for length, was obtained using the formula R−3W. Since this test is speeded, the value is probably an overestimate. The coefficient is based on a sample of 234 unclassified aviation students tested in March 1942 at Psychological Research Unit No. 1.

(3) *Test validity.*—Validation results are given in table 16.13.

Evalaution.—Meter Reading proved to have relatively high navigator validity and low, but definite, pilot validity. Reliability is minimum satisfactory. Administration and scoring are comparatively easy.

Table Reading, CP603B

This is Part III of the Quantitative Perception tests.

Description.—Each item requires the examinee to check whether or not the listed square or square root of a number is correct. A table of squares and square roots is provided for this purpose.

(1) *Internal characteristics.*—The directions contain 8 sample items and a table of squares and square roots of the numbers from 25 to 50. Four of the items are already answered correctly, and four are to be answered by the examinee. The sample items are not scored. The test contains a table of squares and square roots of the numbers from 51 to 100, and 48 recorded and scored items.

TABLE 16.13.—*Validation data*[1] *for Meter Reading, CP602B*

Group	Criterion	N_t	p_g	M_g	M_e	SD_t	r_{bis}
Pilots in primary training	Graduation-elimination	[2]2,497	0.73	[3]5.00	[3]4.34	[3]2.02	0.14
Navigators[4]	........do	[5]228	.79	26.0	19.7	10.2	.36
........Do[6]	........do	[7]365	.90	26.0	17.5	10.0	.44
........Do[8]	........do	[9]183	.84	26.2	20.1	10.4	.32
........Do[8]	........do	[9]392	.77	26.5	21.0	10.8	.34
Bombardiers[4]	Record circular error[10]	[7]557	....	.....	.....	.....	[11].05
........Do[4]	........do[10]	[12]194	....	.....	.....	.....	[11]—.04

[1] On these samples, scoring formulas could not be identified.
[2] Group not identified.
[3] Using scaled scores with a mean of 5.00 and a standard deviation of 2.00.
[4] New students.
[5] In classes 42-10 through 42-17. Tested at Psychological Research Unit No. 1.
[6] New students and eliminated pilots.
[7] Class not identified. Tested from Apr. 1 to Aug. 14, 1942 at Psychological Research Unit No. 3.
[8] Eliminated pilots.
[9] Class not identified. Tested in February, March and April, 1942 at Psychological Research Unit No. 2.
[10] A very unreliable criterion.
[11] Product-moment correlation.
[12] In classes 43-12 through 43-16. Tested at Psychological Research Unit No. 1.

703130—47—26

(2) *Administration.*—Each examinee receives a special IBM form on which both directions and items are presented. Three minutes are allowed to complete the test. Sample items are shown in figure 16.5.

(1) N	SQUARE	R	W	(2) N	SQ. ROOT	R	W
25	625	▬	::::	49	7.071	::::	▬
31	962	::::	▬	37	6.083	▬	::::

TABLE

A NO.	B SQUARE	C SQUARE ROOT
25	625	5.000
26	676	5.099
27	729	5.196
28	784	5.292
29	841	5.385
30	900	5.477
31	961	5.568
32	1 024	5.657
33	1 089	5.7…
34	115…	
35		
3…		

FIGURE 16.5
SAMPLE PROBLEM OF TABLE READING, CP603B

Following is part of the directions:

The • • • table gives the squares and square roots of numbers from 25 to 50. Look up in this table the squares of the numbers in column (1). If the answer given is right, mark the space under R; if it is wrong, mark the space under W. Then look up the square roots of the numbers in column (2) and blacken the answer space under R if the answer given is right or W if it is wrong.

(3) *Scoring.*—When the test was first introduced into the battery, the scoring formula was R−3W. This was changed to R — 2W and finally to R — W.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 16.14.

TABLE 16.14.—*Distribution constants for Table Reading, CP603B*[1]

Group	N	M	SD
Bombardiers[2]	557	26.1	6.9
Navigators[2]	365	28.3	7.2

[1] On these samples scoring formulas could not be determined from available data.
[2] Tested from Apr. 1 to A[illegible] 14, 1942 at Psychological Research Unit No. 3.

(2) *Reliability coefficient.*—By the odd-even method, an estimated reliability coefficient of 0.85, corrected for length, was obtained using the scoring formula R−3W. Since the test is speeded, the coefficient is spuriously high. This figure is based on a sample of 234 unclassified aviation students tested in March 1942 at Psychological Research Unit No. 1.

(3) *Test validity.*—Validation results based on several samples are given in table 16.15.

TABLE 16.15.—*Validity data*[1] *for Table Reading, CP603B*

Group	Criterion	N_t	p_s	M_s	M_e	SD_t	r_{bis}
Pilots in primary training	Graduation-elimination	[2]2,500	0.73	[3]4.73	[3]4.76	[3]1.9	0.00
Navigators[4]	do	[5]228	.79	30.3	26.3	6.3	.37
Do[6]	do	[7]365	.90	28.9	24.7	7.2	.30
Do[8]	do	[4]183	.84	31.3	24.0	7.7	.52
Do[4]	do	[9]163	.71	28.3	24.2	7.1	.36
Do[8]	do	[9]362	.77	27.9	24.6	6.9	.45
Bombardiers[4]	Record circular error[10]	[7]557	....	.....	.....	....	[11].11
Do[4]	do[10]	[12]194	....	.....	.....	....	[11].01

[1] On these samples, scoring formulas could not be identified.
[2] Group not identified.
[3] Using scaled scores with a mean of 5.00 and a standard deviation of 2.00.
[4] New students.
[5] In classes 42-10 through 42-17 tested at Psychological Research Unit No. 1.
[6] New students and eliminated pilots.
[7] Class not identified. Tested from Apr. 1 to Aug. 14, 1942 at Psychological Research Unit No. 3.
[8] Eliminated pilots.
[9] In classes 42-11 to 42-16 at Hondo Army Air Field. Tested February, March and April at Psychological Research Unit No. 2.
[10] A very unreliable criterion.
[11] Product-moment correlation.
[12] In classes 42-12 to 42-16. Tested at Psychological Research Unit No. 1.

Evaluation.—One relatively small samples, Table Reading proved to have moderate to relatively high validity for predicting success in navigator training and, on a very large sample, zero validity for pilots. Test CE603B probably has a satisfactory reliability and is easy to administer and to score.

Number Reading (Filing), CP604B

This is Part IV of the Quantitative Perception tests.

Description.—Each item presents a number to be filed and two alternative numbers after which it might be filed. One master table listing all of the numbers in the file accompanies the items. The examinee is required to determine which of the two alternative numbers the number to be filed should follow.

(1) *Internal characteristics.*—The directions contain a master table and seven sample items, two of which have the correct answers marked and five of which the examinee is required to answer. Sample items are not scored. The test contains one master table and 50 recorded and scored items.

(2) *Administration.*—Each examinee receives a special IBM form in which both directions and items are printed. Three minutes are allowed to complete the test items. Sample items appear in figure 16.6.

(A) NOS. IN FILE
11.8
14.7
20.1
22.2
22.6
31.7
40.8
41.9

(B) NOS. TO BE FILED	TO BE FILED DIRECTLY AFTER	
20.3	14.7	20.1
12.0	11.8	11.9

FIGURE 16.6
SAMPLE PROBLEMS OF NUMBER READING,
CP604B

Following is part of the directions:

Column A is a column of numbers in order of size. In column B are numbers to be filed after the correct numbers in A. The answers are given after each number. Find the correct answers by deciding after which number in A you will file the number you are working on, then blacken the space under the correct answers.

(3) *Scoring.*—When the test was introduced into the classification battery, the scoring formula was R—3W. This was changed to R — 2W and finally to R — W.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 16.16.

TABLE 16.16.—*Distribution constants for Number Reading, CP604B*[1]

Group	N	M	SD
Unclassified aviation students[2]	234	16.6	7.4
Bombardiers[3]	238	15.7	7.4
Navigators[3]	194	18.1	7.0

[1] On these samples scoring formulas could not be determined from available data.
[2] Tested in March 1942 at Psychological Research Unit No. 1.
[3] Tested from Apr. 1 to Aug. 14, 1942 at Psychological Research Unit No. 3.

(2) *Reliability coefficient.*—By the odd-even method, an estimated reliability coefficient of 0.93, corrected for length, was obtained using the scoring formula R−3W. Since the test is speeded, this coefficient is spuriously high. This figure is based on a sample of 234 unclassified aviation students tested in March 1942 at Psychological Research Unit No. 1.

(3) *Test validity.*—Validation results based on several samples are given in table 16.17.

TABLE 16.17.—*Validity data for Number Reading (Filing), CP604B, using the graduation-elimination criterion*[1]

Group	N_t	p_s	M_s	M_e	SD_t	r_{bis}
Navigators[2]	194	0.88	19.4	13.6	7.0	0.44
Do[3]	163	.71	19.8	12.5	7.9	.55
Do[4]	392	.77	19.7	13.8	7.7	.45

[1] On these samples scoring formulas could not be determined from available data.
[2] New students and eliminated pilots, tested from Apr. 1 to Aug. 14, 1942 at Psychological Research Unit No. 3.
[3] New students in classes 42-11 to 42-16 at Hondo Army Air Field tested at Psychological Research Unit No. 2.
[4] Eliminated pilots. Same classes and tested in same unit as in footnote 3.

Evaluation.—On relatively small samples, Number Reading proved to have high navigator validity. The test has satisfactory reliability and is easy to administer and to score.

Number Size, CP605B

This is part V of the Quantitative Perception tests.

Description.—The examinee's task in this test is to scan rapidly rows of numbers and to underline those that fall within a specified number range.

(1) *Internal characteristics.*—The test is divided into five sections. Each section contains 75 numbers ranging from 1 to 99. There are 70 numbers to be underlined, making a total possible score of 70, if all the correct numbers are underlined and all of the incorrect numbers are not.

(2) *Administration.*—Each examinee receives a special IBM form on which both directions and items are printed. Two minutes are allowed to complete the test items. Sample problems from the test are shown in figure 16.7.

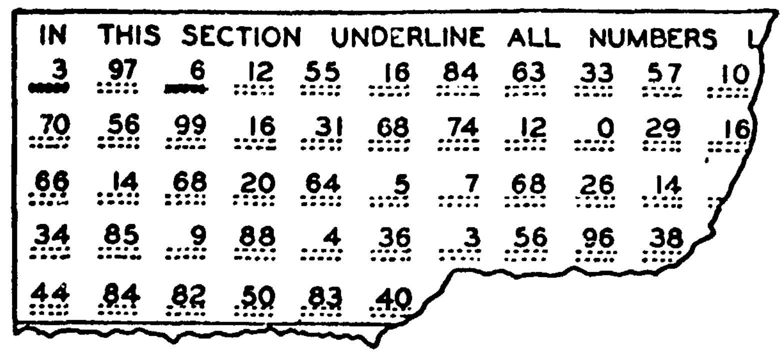


FIGURE 16.7
SAMPLE PROBLEMS OF NUMBER SIZE,
CP605B

Following is part of the directions:

In each of the sections the answer spaces under certain numbers are to be blackened. Follow the directions given at the beginning of each section. Work as rapidly as you can without making mistakes; an error will deduct three points from your score.

(3) *Scoring.*—The scoring formula used is R−3W.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 16.18.

TABLE 16.18.—*Distribution constants for Number Size, CP605A*

Group	N	M	SD
Bombardiers[1]	238	37.5	8.9
Navigators[1]	194	39.8	7.4

[1] Tested from Apr. 1 to Aug. 14, 1942 at Psychological Research Unit No. 3.

(2) *Test validity.*—Validation results based on several samples are given in table 16.19.

TABLE 16.19.—*Validity data for Number Size, CP605B using the graduation-elimination criterion*

Group	N	p_s	M_s	M_e	SD_t	r_{bis}
Pilots in primary training[1]	676	0.60	39.0	37.8	7.1	0.10
Navigators[2]	194	.88	40.8	35.1	7.4	.41
Do[3]	163	.71	39.0	36.7	8.7	.16
Do[4]	392	.77	40.3	37.8	8.0	.18

[1] In classes 42H and 42I.
[2] New students and eliminated pilots. Tested from Apr. 1 to Aug. 14, 1942 at Psychological Research Unit No. 3.
[3] New students. In classes 42-11 to 42-16 at Hondo Army Air Field. Tested at Psychological Research Unit No. 2.
[4] Eliminated pilots. Classes and testing unit same as in footnote 3.

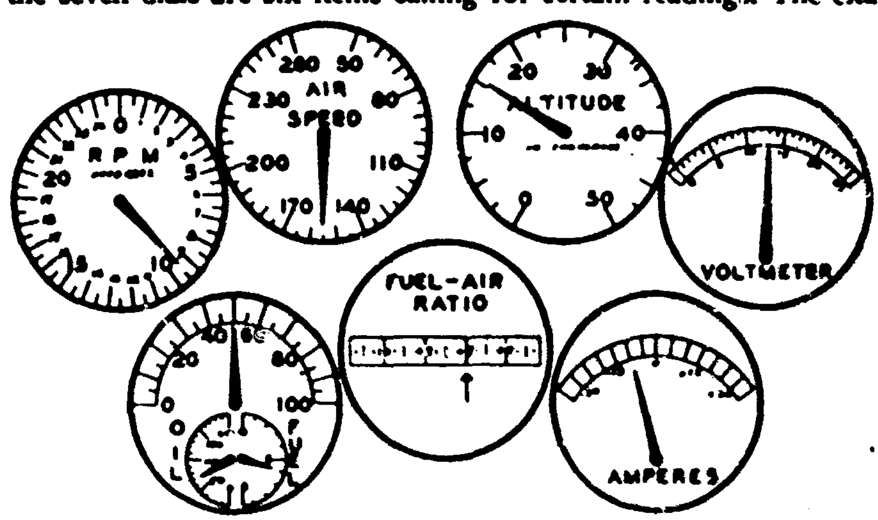
Evaluation.—Number size showed a moderate validity for navigators and a very low validity for pilots.

Dial Reading and Table Reading, CP622A and CP621A *

In some phase of their work, pilots, bombardiers, and navigators all find it necessary to observe dials and to consult printed tables. An activity thus common to three air-crew specialties could hardly escape early notice in the job analyses. Moreover, dial reading and table reading are activities that lend themselves very readily to printed-test presentation. As was true of most early tests whose underlying content had yet to be revealed by factor analysis, the rationale for the adoption of dial-and table-reading tests was the common sense one of merely expecting validity commensurate with the extent to which air-crew members depend upon an activity such as that measured by this test.

Description.—As suggested by the title and the code numbers, the test booklet is a composite of two tests which once were treated as separate units.

CP622A, the first section of the booklet, is the Dial Reading test, occupying the first eight pages. On a typical page a bank of seven dials is drawn on the upper half and repeated with different needle positions on the lower half of the page. The dials are labeled: RPM, Airspeed, Altitude, Voltmeter, Temperature, Fuel-Air Ratio, Amperes. The dials differ widely in graduation ranging from 2,500 units portrayed on the RPM dial down to only five units on the Fuel-Air Ratio dial. Indicator needles point to given values on the various dials. Below each printing of the seven dials are six items calling for certain readings. The exami-



SAMPLE PROBLEMS:	A	B	C	D	E
1. R.P.M.	91.	89.5	9.5	92.	105.
2. AMPERES	16.	-15.5	14.	-10.	-2.
3. ALTITUDE	157.5	15.5	155	1.5	152

FIGURE 16.8
SAMPLE PROBLEMS OF DIAL READING,
CP622A

* Developed at the Office of The Air Surgeon, Headquarters, Army Air Forces. Chief contributor: Lt. Frank J. Dudek. The account of this test was written by Lt. John W. Howe, Jr.

nee's task is first to look at the item to see which dial is to be read; next, to locate the dial and read it correctly (interpolating if necessary); then to identify this reading among the five choices offered; and, finally, to mark this choice on the separate answer sheet. In figure 16.8 are shown the dials and sample problems.

CP621A, the Table Reading test, is divided into two parts. Part I makes use of a large rectangular table of 35 columns and 35 rows, presenting respectively first values and second values. The center column and row are labeled O. Negative values are listed in the columns to the

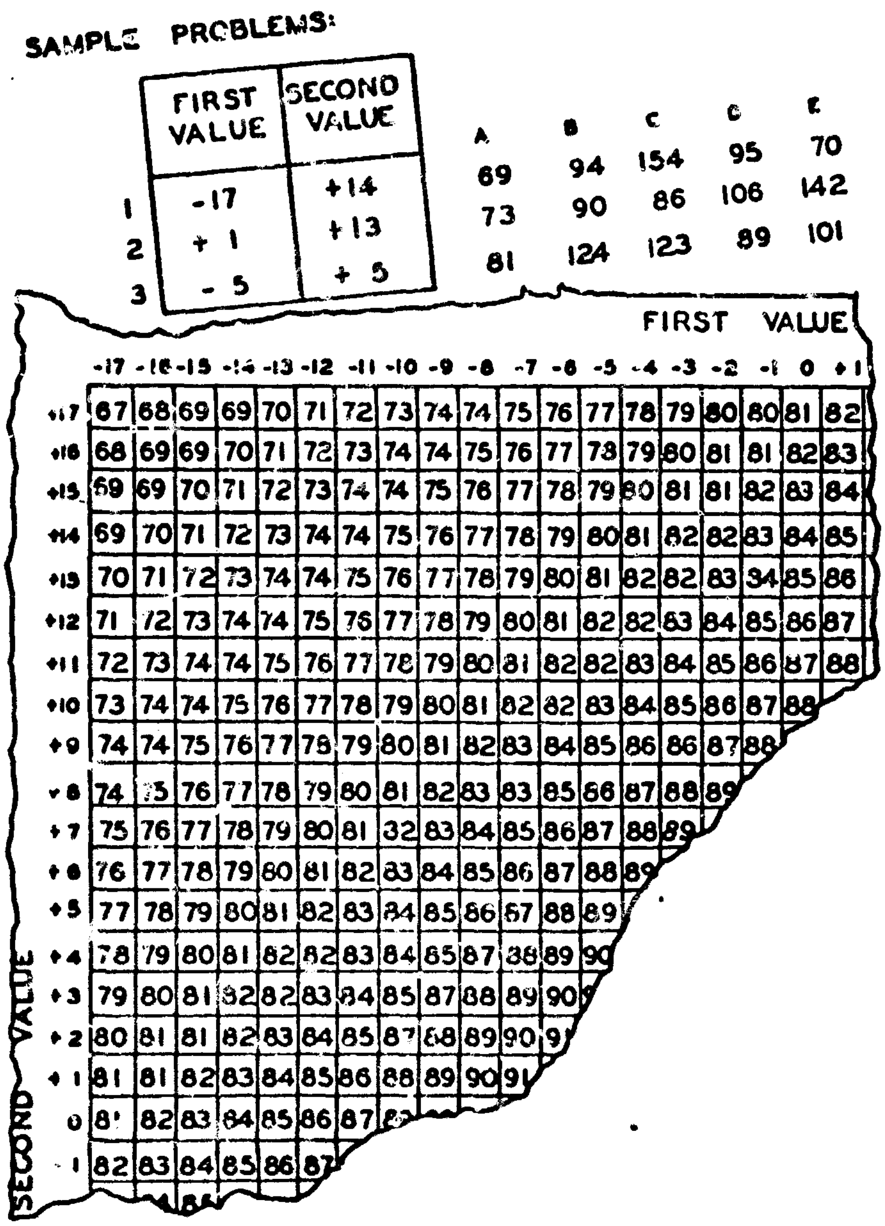


FIGURE 16.9
SAMPLE PROBLEMS OF TABLE READING,
CP621A - PART I

left of center and the rows below the center, while positive values read upwards and to the right of center. Each test item lists two values; one first value and one second value. The examinee's task is to locate the correct column and correct row, read the entry at their intersection, identify it among the five choices listed for the item, and mark the choice on his answer sheet. In figure 16.9 is the table and sample problems.

Part II of CP621A uses four tables of fairly complex construction, containing five variables associated with the flight of an airplane. The five variables are air speed, angle of wind, velocity of the wind, correction for drift with the wind, and ground speed. The first three are considered the independent variables by which the last two are determined. Thus, in each of the 45 items, the air speed, the wind angle, and the wind velocity are given, and the examinee is required to find either the drift correction or the ground speed. Four specific air speeds determine which of the four tables should be entered; wind velocity determines the column; and wind angle determines the row. Figure 16.10 shows a sample item and table.

(1) *Internal characteristics.*—The directions for dial reading contain three sample items with the correct answers marked and three recorded,

AIR SPEED 100 MILES PER HOUR

	WIND VELOCITY					
	10 MPH		15 MPH		20 MPH	
WIND ANGLE	DIR. COR.	GRO. SPD.	DIR. COR.	GRO. SPD.	DIR. COR.	GRO. SPD.
0	0	90	0	85	0	80
10	1	90	1	85	2	80
20	2	90	3	86	4	
30	3	91	4	87	6	
40	4	92	6	88		
50	4	93	7			
60	5	94				
70	5	96				
80	6	98				
90	6					

SAMPLE PROBLEMS:

	AIR SPEED	WIND VELO.	WIND ANGLE	WHAT IS THE	A	B	C	D	E
1	100	20	10	DIR. COR.	6	9	5	2	8
2	100	10	50	GRO. SPD.	93	98	87	85	90

FIGURE 16.10
SAMPLE PROBLEMS OF TABLE READING,
CP621A --- PART II

but unscored, practice items. Dial reading contains 57 recorded and scored items.

The directions for Table Reading, part I, contain three sample items with the correct answers marked and two recorded, but unscored, practice items. Table Reading, part I, contains 43 recorded and scored items. The directions for table reading, part II, contain two sample items with the correct answers illustrated and two recorded, but unscored, sample items. Table Reading, part II, contains 43 recorded and scored items.

(2) *Administration.*—As previously mentioned, CP622A and the two parts of CP621A are all printed in the same booklet. The 3 units are administered in succession, the time limits being 9, 8, and 7 minutes respectively. Answers for all three units are entered on a single answer sheet. Sample problems and all necessary directions are printed in the test booklet.

(3) *Scoring.*—For the 5 months following their insertion in the battery, both tests were scored by the formula R−W. During this period, the two tests were treated as discrete measures, separate scores being secured for each. Statistical study of these scores revealed, however, that the two tests are functionally very similar, so that, since that time, they were scored as one test, using the formula (R−W)/2.

Statistical results.—Since this test was included in the classification battery, it was extensively analyzed.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 16.20.

TABLE 16.20.—*Distribution statistics for Dial and Table Reading, CP622A and CP621A, based upon samples of unclassified aviation students*

Form	Psychological research unit No.	Testing dates	N	M	SD
CP622A	1	October 1942	1,454	22.1	7.5
CP622A	1	Do	1,980	21.7	7.4
CP622A	2	August and September 1942	1,520	24.0	7.4
CP622A	3	March 1943	392	25.3	7.2
CP621A	1	October 1942	1,455	39.6	13.8
CP621A	1	Do	1,991	38.1	13.5
CP621A	2	August and September 1942	1,520	39.7	14.1
CP621A	2	October 1942	2,376	39.4	14.6
CP621A	3	March 1943	392	43.1	12.2
CP621A and CP622A	1	December 1942	1,096	32.0	9.1
CP621A and CP622A	2	Do	1,015	34.0	10.0
CP621A and CP622A	3	Do	1,143	33.7	9.0
CP621A and CP622A	1, 2, 3	July 1943	3,000	33.9	8.8
CP621A and CP622A	1, 2, 3	November 1943	1,500	34.5	8.9
CP621A and CP622A	[1]4–10	Do	1,920	32.4	9.8

[1] Medical and Psychological Examining Units.

(2) *Internal consistency.*—The degree of homogeneity of the items, of all three parts combined, is indicated by a mean internal-consistency phi of 0.20, a standard deviation of the phi distribution of 0.07, and a range of values from 0.04 to 0.42. These statistics are based upon the responses of the highest 25 percent and the lowest 25 percent in total

score of a group of 800 unclassified aviation students tested in October 1943 at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—Five samples yielded the estimates of reliability given in table 16.21.

TABLE 16.21.—*Estimated reliability coefficients of Dial Reading Test, CP622A, and Table Reading Test, CP621A*

Form	Group	Type	N	r'_{tt}	r_{tt}
CP622A	Unclassified aviation students[1]	Alternate-forms[2]	1,167	0.62	0.76
CP621A	Do[1]	Alternate-forms[2]	1,167	.73	.84
CP621A and 622A	Do[1]	Alternate-forms[2]	1,167	.77	.87
CP621A and 622A	Do	Test-retest[3]	712	....	.82
CP621A and 622A	Do[4]	Equivalent halves	1,000	.81	.90

[1] Tested at Medical and Psychological Examining Unit No. 8. Testing dates not reported; date reported May 1945.
[2] Administered for experimental purposes in two separately timed halves.
[3] 28-day interval between testing. Tested at Medical and Psychological Examining Unit No. 6 from 11 to 19, April 1945.
[4] Tested at Medical and Psychological Examining Unit No. 7, April 1944.

(4) *Difficulty.*—Based upon the responses of the above-mentioned sample of 800 unclassified aviation students, the three parts of the test yielded a mean proportion of correct responses of 0.85, corrected for chance, with a range from 0.25 to 0.99 and a standard deviation of 0.09.

(5) *Factorial composition.*—For all parts combined the most significant loadings are in the numerical (0.53), space I (0.42), and perceptual-speed (0.31) factors. The communality is 0.65. For a full picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Validation results are given in tables 16.22 to 16.25 inclusive.

(7) *Item validity.*—Validation of items of this test disclosed the results recorded in table 16.26.

Evaluation.—As indicated by the code number, the test was primarily thought to be a test of perception. On a priori grounds this assumption was entirely reasonable. One would certainly expect that rapidly looking through tables and inspecting dials would draw markedly upon perceptual ability. In part, this did prove to be the case, for 11 percent of the total variance of the test is accounted for by the perceptual-speed factor. In addition, however, two other factors show even greater saturation in the test. Twenty-eight percent of the total variance is numerical and 19 percent is in spatial relations. The communality represents 65 percent of the total variance. Thus, reading a set of dials and tables involves more than might be expected at first glance. Apparently, numerical ability is required in the implicit additions and subtraction while interpolating dial readings. In table reading, the use of positive and negative numbers and of quadrants of the table is no doubt a contributing influence to the numerical loading.

TABLE 16.22.—*Validity data for Dial Reading, CP622A, and Table Reading, CP621A, for pilots in training, using the graduation-elimination criterion*

Form	Group	Psychological research unit No.	Class	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$
CP622A	In primary training	2	43D	1,520	0.75	24.7	22.0	7.4	0.22	.....
CP622A	Do	2	43I	1,148	.76	23.6	22.4	7.2	.10	.....
CP622A	Do	2	43F	2,376	.63	24.2	22.4	7.4	.15	.....
CP622A	In basic training	2	43F	1,429	.84	24.5	22.5	7.4	.15	.....
CP621A	In primary training	2	43D	1,520	.75	41.0	35.9	14.1	.22	.....
CP621A	Do	2	43I	1,148	.76	38.7	36.2	13.8	.11	.....
CP621A	Do	2	43F	2,376	.63	40.9	37.0	14.6	.16	.....
CP621A	In basic training	2	43F	1,429	.84	41.3	38.7	14.1	.10	.....
CP622A, CP621A	In primary training	2	43H	3,038	.66	68.5	61.7	19.5	.22	.....
CP622A, CP621A	Do	1, 2, 3	43J	3,151	.66	33.3	31.2	8.8	.15	[1]0.20
CP622A, CP621A	Through basic training	1, 2, 3	43J	3,046	.57	33.4	31.5	8.8	.14	[1].19
CP622A, CP621A	Through advanced training	1, 2, 3	43J	2,978	.54	33.5	31.4	8.8	.15	[1].20
CP622A, CP621A	In primary training	1, 2, 3	44E	4,779	.88	35.8	32.3	8.5	.21	[1].23
CP622A, CP621A	Do	1, 2, 3	44I	3,145	.84	35.2	32.2	7.9	.21	[3].26
CP622A, CP621A	In B–17 transition training	1, 2, 3	43J–44A	1,046	.98	34.6	33.0	7.9	.08	.....
CP622A, CP621A	In B–24 transition training	1, 2, 3	43J–44A	982	.92	34.5	31.0	8.7	.20	.....
CP622A, CP621A	In B–25 transition training	1, 2, 3	43J–44A	314	.96	33.6	35.0	8.4	−.07	.....
CP622A, CP621A	In B–26 transition training	1, 2, 3	43J–44A	380	.82	35.6	33.0	8.4	.17	.....
CP622A, CP621A	In P–40 assignment	1, 2	43J–44A	2,416	.64	34.4	33.5	8.5	.07	.....
CP622A, CP621A	WASPS[2]	......	44–W–8	104	.61	36.0	29.8	8.6	.45	.....

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Women's auxiliary service pilots.
[3] Assuming an unrestricted stanine standard deviation of 1.90.

TABLE 16.23.—*Validity data for Dial Reading, CP622A, and Table Reading, CP621A, for navigators and bombardiers in training, using the graduation-elimination criterion*

Form	Group	Psychological research unit No.	Class	N_t	r_t	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
CP622A, CP621A	Navigators	1, 2	43–10, 43–11	731	0.87	39.2	31.9	8.2	0.48	0.60
CP622A, CP621A	Do	1, 2, 3	43–12 through 43–15	1,968	.79	38.2	32.4	8.1	.41	.54
CP622A, CP621A	Bombardiers[2]	1, 2, 3	43–8 to 43–11	1,829	.79	34.1	31.7	8.8	.15	.18
CP622A, CP621A	Do[3]	1, 2, 3	43–14 to 43–18	455	.84	33.4	28.9	8.8	.28	.31
CP622A, CP621A	Do[4]	1, 2, 3	43–14 to 43–18	524	.85	37.4	35.9	7.7	.11	.17
CP622A	Do[2]	1	43–5 to 43–7	552	.84	19.0	15.6	7.3	.25	....
CP622A	Do[2]	2	43–5 to 43–7	330	.86	24.5	23.2	8.3	.09	....
CP622A	Do[2]	3	43–5 to 43–7	496	.82	25.1	23.1	7.4	.16	....
CP621A	Do[2]	1	43–5 to 43–7	552	.84	34.7	29.8	12.7	.21	....
CP621A	Do[2]	2	43–5 to 43–7	329	.86	42.8	34.7	15.3	.28	....
CP621A	Do[2]	3	43–5 to 43–7	469	.83	44.3	38.2	12.5	.27	....

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] New aviation students, taking the 12-week course (no navigation training).
[3] New aviation students taking the 15-week course (some navigation training).
[4] Reclassified pilots, taking the 18-week course.

TABLE 16.24.—*Validity data for Dial and Table Reading, CP621A–CP622A, using seven grades in navigation training as criteria, for a sample of 463 navigation trainees in Hondo, classes 43–10 through 43–15*

Grade	r[1]	r_{corr}[2]
Dead reckoning (ground school)	0.39	0.52
Celestial navigation (ground school)	.35	.45
Dead reckoning (flight)	.26	.33
Celestial navigation (flight)	.31	.41
Meteorology	.15	.29
Military	.11	.17
Final composite	.41	.54

[1] Product-moment correlation.
[2] Assumed unrestricted stanine standard deviation not reported.

TABLE 16.25.—*Validity data for Dial Reading, CP622A, and Table Reading, CP621A, using various criteria*

Form	Group	Class	Psychological research unit No.	Criterion	N	r[1]
CP622A	Bombardiers in training	43-1 to 43-4	3	Academic averages	195	0.27
CP622A	Do	43-1 to 43-4	3	Record circular error	195	−.07
CP622A	Do	43-1 to 43-4	3	do	195	−.13
CP622A	Flexible gunners in training	43-27 to 43-30	1	Air-to-air firing	311	.03
CP622A	Do	43-27 to 43-30	1	Academic averages	311	.21
CP621A	Bombardiers in training	43-1 to 43-4	3	do	195	.27
CP621A	Do	43-1 to 43-4	3	Record circular error	195	−.03
CP621A	Do	43-1 to 43-4	3	Combat circular error	195	−.15
CP621A	Flexible gunners in training	43-27 to 43-30	1	Air-to-air firing	250	−.06
CP621A	Do	43-27 to 43-30	1	Academic averages	250	.12
CP622A, CP621A	Pilots in training	43J	1	Gunnery: Total percent hits	459	.11
CP622A, CP621A	Do	43J	1	Gunnery: Qualification percent hits	459	.02
CP622A, CP621A	Do	43K	1	Gunnery: Total percent hits	403	.00
CP622A, CP621A	Do	43K	1	Gunnery: Qualification percent hits	403	.02
CP622A, CP621A	Pilots in advanced single-engine training	43J—44A	3	Proficiency rating	562	.14
CP622A, CP621A	Pilots in advanced twin-engine training	43J—44A	3	do	685	.09
CP622A, CP621A	Pilots in four-engine transition training	43J—44A	3	do	685	.05
CP622A, CP621A	Air mechanics in training		1, 2, 3	Average grades	232	.32
CP622A, CP621A	Armorers in training	34-44A—35-44A	[2]1-10	do	269	.31
CP622A, CP621A	Do		1, 2, 3	do	376	.05
CP622A, CP621A	Radio-operator mechanic in training		1, 2, 3	Graduation-elimination	235	.18
CP622A, CP621A	Radio-operator mechanic in training		1, 2, 3	Average grades	153	.27
CP622A, CP621A	Officer candidates		[2]5	5th week academic average	343	.33
CP622A, CP621A	Do		[2]5	8th week ratings	343	.17
CP622A, CP621A	Do		[2]5	12th week ratings	343	.15

[1] Product-moment correlation.
[2] Including Medical and Psychological Examining Units.

TABLE 16.26.—*Validity of items of Dial Reading, CP622A, and Table Reading, CP621A, using the graduation-elimination criterion*

Group	N	r_e	Mϕ	SDϕ	Range of ϕ	
					Low	High
Pilots in primary training[1]	1,400	0.57	0.02	0.03	−0.07	0.14
Navigators[2]	746	.81	.04	.05	−.07	.16

[1] In class 43I. Tested in March and April 1943 at Psychological Research Unit No. 3.
[2] Class not identified. Tested in April and May 1944 at Psychological Research Unit No. 3.

The Dial and Table Reading test attained its highest validity in predicting navigator success. This it did unusually well, the corrected coefficient for the sample of nearly 2,000 cases being 0.54. For success in primary pilot training, the test's predictive power was much more moderate. The coefficients range mostly from 0.20 to 0.28.

Of all the tests in the classification battery, Dial and Table Reading has the distinction of being the best single predictor of success for any aircrew specialty. It predicts navigator success with a validity of approximately 0.54. The explanation, no doubt, lies in the fact that two important fundamental navigator abilities are measured by the test to a substantial degree. These are the numerical and the spatial-relations factors. Speed of perception is also sampled, and it contributes a small but appreciable amount to the navigator validity. These three factors account for about 80 percent of the obtained validity, and small loadings on reasoning I, mathematical background, verbal, and other factors very nearly account for the remaining 20 percent. The pilot validity is likewise fully accounted for by familiar factors.

It should be pointed out, however, that the three leading factors contained in Dial Reading and Table Reading are better measured by other battery tests. In spite of the high navigator validity, therefore, it cannot be considered that the prediction of navigators is the greatest value of the test, for that can be done on the basis of other tests. Dial and Table Reading did aid in the discovery and definition of these factors, however, and served to give secondary and supplementary coverage of them in the classification battery.

Paratroop Dropping Test, CI209A [7]

This test was designed to measure table reading and computational ability in a meaningful situation.

Description.—Each item of the test requires the examinee to select from five diagrammatic sketches the one that portrays a situation in which a paratrooper would land directly on his objective.

Each of the diagrammatic sketches shows an airplane in flight, having passed over and beyond its objective, an enemy city. The distance beyond the objective is given. The task of the examinee is to determine whether or not the given distance is equal to the distance that the

[7] Developed at Medical and Psychological Examining Unit No. 10. Chief contributor: Capt. Joseph E. King.

parachutist would land (considering plane height, air speed, wind speed, and diameter of the parachute) behind the airplane. If the two distances are equal, the examinee concludes that the paratrooper would land directly upon his objective.

(1) *Internal characteristics.*—The directions contain three recorded, but unscored, sample items. The test contains 22 scored items.

(2) *Administration.*—Ten minutes are allowed for the directions and 30 minutes for the test problems.

Following are parts of the directions:

A pilot is assigned the task of dropping paratroops upon an enemy town. In dropping the chutists he must consider four factors; (1) The height of the airplane from the ground, (2) the diameter of the parachute, (3) the speed of the wind against which he is flying, and (4) the speed of the airplane.

The effect of each factor is stated in terms of the feet behind the airplane that the chutist will land. Two or more factors may operate at the same time, as, for example, speed of the wind and diameter of the parachute. When this happens, the chutist lands a distance behind the plane which is the sum of the two or more distances.

Distances behind the airplane that the chutist will land, under the various conditions imposed, are given in tables accompanying the problems.

(3) *Scoring.*—The scoring formula used is R−W/4.

Statistical results.—The limited data available are for examinees tested at Medical and Psychological Examining Unit No. 10.

(1) *Distribution statistics.*—A sample of 400 unclassified aviation students yielded a mean score of 14.9 and a standard deviation of 5.1.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a median internal-consistency phi of 0.60.

(3) *Reliability coefficient.*—By the odd-even method, an estimated reliability coefficient of 0.73, uncorrected, was obtained. This figure is based on a sample of 400 unclassified aviation students.

Evaluation.—Correlations with battery tests indicated that the Paratroop Dropping test was not measuring any factors not already covered by other tests, so validation was not recommended.

Marking Accuracy (no code number)[a]

This test was constructed as part of an experiment designed to determine whether test validities and other correlations were affected by speed in marking answer sheets. All printed tests are machine scored, answers being recorded on a separate answer sheet. This operation is a clerical function which, in speed tests, may influence the scores to a

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Frank Dudek and Capt. Lloyd G. Humphreys.

practical extent. If the influence of any irrelevant function proves to be real, it can be subtracted, permitting the intended nature of each test to be more prominent.

Description.—The test material consists of IBM answer sheets on which the letters designating answer positions are circled by overprinting. The examinee's task is merely to blacken the space under the circled letters.

(1) *Internal characteristics.*—The test is divided into 2 parts, each containing 75 items.

(2) *Administration.*—The directions are delivered orally. Forty seconds are allowed to mark the items in each part. The total testing time is approximately three minutes. Following is part of the orally-administered directions:

We are interested in determining how much of a dexterity factor there is in a paper-and-pencil test. For this reason, we are asking you merely to fill out an answer sheet as rapidly as you can.

You have an answer sheet on which small circles indicate the spaces to be blackened. It is important that the correct spaces be blackened and that they be blackened adequately enough to be scored on an electric scoring machine.

(3) *Scoring.*—The score is simply the number of the last marked item.

Statistical results.—The available data are based upon examinees tested at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—A sample of 284 classified pilots tested in March and April 1942 and August and September 1943 yielded a mean score of 83.6 and a standard deviation of 8.0. The distribution curve is slightly positively skewed and somewhat flatter than normal.

(2) *Reliability coefficient.*—By the alternate-forms (part I-part II) method, an estimated reliability coefficient of 0.86, corrected for length, was obtained. This figure is based on the above-mentioned sample of 284 classified pilots.

(3) *Factorial composition.*—The only substantial loadings are in the psychomotor III (0.50) and perceptual-speed (0.35) factors. The communality is 0.41. For a fuller picture of the factorial composition of this test, see appendix B.

Evaluation.—Marking accuracy, like Log Book Accuracy (described below), was factor-analyzed in the Integration Battery (see ch. 10). Marking Accuracy differs from Log Book Accuracy only in its secondary loadings. Both tests are highly saturated with the psychomotor-speed factor. The Marking Accuracy test shows some perceptual-speed variance (loading of 0.35), in contrast to an equivalent loading on the numerical factor for Log Book Accuracy. Apparently the task of locating the reference circles on the answer sheet demands a substantial amount of perceptual-speed ability.

703320—47—27

Log Book Accuracy (no code number)[*]

This test was also designed for the purpose of determining the part that speed of marking an answer sheet might play in many printed tests.

Description.—In the test booklet are listed item numbers followed by the answers A, B, C, D, or E to be marked. The item numbers are in random order rather than in sequence. The examinee's task is to record the answers quickly and correctly on a separate answer sheet.

(1) *Internal characteristics.*—The directions contain an illustration of five correctly marked items. Parts I and II each contain 75 recorded and scored items.

(2) *Administration.*—Four minutes are allowed for each part, while directions consume about 3 minutes, making a total testing time of approximately 11 minutes.

Following are parts of the directions and sample items:

Opposite each of the item numbers in the test booklet is a letter. Your only task is to blacken the space on your answer sheet which corresponds to the item number and letter in the booklet.

Look at the five sample items:

1. C.
4. A.
3. D.
2. E.
5. B.

If you were marking your answer sheet, you would blacken space C opposite item 1; space A opposite item 4; and so on.

(3) *Scoring.*—The scoring formula is R−W.

Statistical results. (1) *Distribution statistics.*—A sample of 278 classified pilots tested at Psychological Research Unit No. 3 in August and September 1943 yielded a mean score of 66.2, and a standard deviation of 11.0. The distribution curve is approximately symmetrical and considerably flatter than normal.

(2) *Reliability coefficient.*—By the alternate-forms (part I-part II) method, an estimated reliability coefficient of 0.75, corrected for length, was obtained. This figure is based on the above-mentioned sample of 278 classified pilots.

(3) *Factorial composition.*—The only substantial loadings are in the psychomotor III (0.60) and numerical (0.32) factors. The communality is 0.56. For a fuller picture of the factorial composition of this test, see appendix B.

Evaluation.—The only available data on Log Book Accuracy are in connection with the factor analysis of integration tests (see ch. 10). The

[*] Developed at Psychological Research Unit No. 3, Chief contributors: Capt. Stuart W. Cook and Lt. Frank J. Dudek.

two marking tests, Log Book Accuracy with a loading of 0.60 and Markin Accurary with a loading of 0.50, defined a new factor hitherto unknown. Whether the factor is broader than speed of marking cannot be answered from the available data. A loading of 0.32 for Log Book Accuracy on the numerical factor indicates that the task of locating answer-sheet numbers quickly when they are out of sequence involves numerical facility. Some spatial ability is possibly involved here also, but the loading in the spatial-relations factor (0.19) is hardly significant enough to support adequately this assumption.

The principal value of Log Book Accuracy and Marking Accuracy was to show that if marking speed affects the scores on the majority of machine-scored tests, it does so only slightly. In no test included in the integration analysis, many of which have complicated marking directions, did marking speed constitute a serious source of extraneous variance, for in none is there an appreciable loading with this factor.

Evaluation of Clerical Speed Tests

This group of tests was shown to measure abilities that are important to air-crew members, especially the navigator, in achieving success in training. The best representative measure of the group, Dial and Table Reading, has occupied a permanent place in the classification battery since the early months of testing.

The principal features of these tests are their numerical and perceptual-speed content. In some of these tests spatial-relations factorial content is also present. The numerical factor is the most valid factor for navigators yet measured, while perceptual speed and spatial relations are valid for all three air-crew specialties.

Although there was no special intent by the AAF to analyze the nature of clerical tasks as such, the similarities of the problems presented by this group of tests offers an opportunity to speculate regarding clerical activities in general. Since they seem to be primarily perceptual and numerical, a test battery composed of pure measures of these two factors weighted properly might be adequate for measuring aptitude for many types of clerical work. For both factors, relatively pure measures are available. There is some reason to believe that two other factors may also enter into aptitude for certain types of clerical work. These are psychomotor speed and spatial relations. In Log Book Accuracy we have a fair measure of psychomotor speed, but we need more data upon which to base an interpretation of the factor before we can safely prescribe its use. For spatial relations, no independent measures have yet been discovered, although there are several tests with significant saturations in it. Further discussions of this trait will be found in later chapters.

EVALUATION OF PERCEPTUAL SPEED TESTS

Factor analysis of experimental and classification tests developed for use in the selection of air-crew members has, so far, revealed only one factor to which the name perceptual has been applied. This factor has been described as the ability to discriminate rapidly visual differences in form and detail. The factor is best defined by Speed of Identification, a test similar to Thurstone's Identical Forms test which, in his original analysis of 56 variables, best defined the factor P.

Tests that reveal relatively pure or independent factors were of particular value to aviation psychologists. They were very useful, for example, in factor-analysis studies. In setting up a correlational matrix to be analyzed, factor analysts recognize that the inclusion of such tests helps to simplify the rotational procedure, particularly in determining the initial directions of rotation. Lesser known factors can then be revealed more quickly and clearly. The greatest value of pure tests is their adaptability in selection testing. If a job is evaluated in terms of factors, it is necessary only to procure an independent measure of each factor involved and to weight each measure properly in the final evaluation.

The fact that perceptual speed is involved to a greater or lesser extent in such a variety of jobs and occupations insures a permanent use for a pure measure of the factor. Tests such as Speed of Identification and Identical Forms are thus of considerable value because they are among the purest measures of single known factors.

FACTOR ANALYSES OF PERCEPTUAL TESTS

The Data

Two factor analyses were made of perceptual tests. A battery of 31 tests was administered to 392 unclassified aviation students. The battery consisted of a number of tests designed to measure different aspects of perception plus selected classification-battery tests of known factorial content. For the purposes of analysis, the battery was divided into two smaller groups, 18 tests in the first matrix and 22 in the second, with 12 tests common to both. While there are disadvantages in making two smaller analyses in place of a single large one, it was believed that the greater ease of computation and simplicity of rotation would make the former method more profitable. The correlation matrices are given in tables 16.27 and 16.28.[10]

All the tests included in these analyses are described elsewhere in this volume, with the exception of the four that are briefly described in the following paragraphs. Full descriptions of these tests can be found in reports 6 and 7 of the AAF psychological series.

Speed Estimation II Identification of Velocities, CP205B—II.—This is a motion-picture test presenting a model plane against a moving background (clouds). The airplane shown in the middle of the screen thus appears to be moving. The examinee is taught to recognize five different

[10] We are highly indebted to Professor L. L. Thurstone who kindly placed at the disposal of the AAF a number of his perceptual tests.

TABLE 16.27.—*Correlation matrix for Perceptual Battery I (N=392)*[1]

Test	1	2	3	4	5	6	7	8	10	13	15	17	20	21	23	24	26	31
1. Speed of Identification (R)[2]	...	384	332	236	295	016	126	295	315	064	210	543	239	310	340	267	248	232
2. Spatial Orientation II	384	.....	237	209	172	−031	361	309	246	121	293	338	224	331	234	280	218	265
3. Spatial Orientation I	332	237	...	333	382	387	073	138	271	035	287	279	237	218	284	343	352	421
4. Dial Reading	236	209	333	...	525	544	150	206	307	082	432	164	371	254	327	312	204	316
5. Table Reading	295	172	382	525	.....	378	−001	359	292	031	324	251	313	207	406	276	233	301
6. Mathematics B	016	−031	387	544	378	.....	182	042	179	152	439	019	327	151	124	252	298	274
7. Mechanical Principles	126	361	073	150	−001	182	.....	270	309	289	287	043	300	267	−052	190	256	145
8. SAM Complex Coordination	295	309	138	206	359	042	270	...	256	202	308	204	257	321	252	236	183	190
10. Path Distance	315	346	271	307	292	179	309	256	...	149	282	258	314	292	259	266	232	242
13. Map Distance	064	121	035	082	031	152	289	202	149	...	204	022	199	141	028	189	235	158
15. Directional Orientation B	210	293	287	432	324	439	287	308	282	204	...	114	411	376	169	367	305	394
17. Speed of Identification (N-R)	543	338	279	164	251	019	043	204	258	022	114	...	205	303	348	244	195	180
20. Cubes	239	224	237	371	313	327	300	257	314	199	411	205	...	459	194	265	267	231
21. Flags, Figures, Cards	310	331	218	254	207	151	267	321	292	141	376	303	459	...	236	324	207	250
23. Path Tracing	340	234	284	327	406	124	−052	252	259	028	169	348	194	236	.....	343	166	252
24. Pattern Analysis	267	280	343	312	276	252	190	236	266	189	367	244	265	324	343	...	282	361
26. Gottschaldt Figures, Part III	248	218	352	204	233	298	256	183	232	235	305	195	267	207	166	282	...	318
31. Plane Formation	232	265	421	316	301	274	145	190	242	158	394	180	231	250	252	361	318	...

[1] Decimal points omitted.
[2] For code numbers see table 16.29.

TABLE 16.28.—*Correlation matrix for Perceptual Battery II (N=392)*[1]

Test	1	5	6	7	8	9	11	12	13	14	15	18	19	20	21	22	24	26	27	29	31
1. Speed of Identification (R)[2]	..	30	02	13	30	19	10	26	06	15	21	32	11	24	31	30	27	25	34	15	23
5. Table Reading	30	..	38	00	36	16	20	06	03	12	32	28	17	31	21	24	28	23	31	09	30
6. Mathematics	02	38	..	18	04	10	19	05	15	14	[illegible]	36	14	33	15	18	25	30	22	−06	27
7. Mechanical Principles	13	00	18	..	27	19	03	14	29	22	29	35	10	30	27	30	19	26	18	09	14
8. SAM Complex Coordination	30	36	04	27	..	19	10	13	20	15	31	28	12	26	32	30	24	18	26	22	10
9. Point Distance	19	16	10	19	19	..	24	18	24	19	26	28	13	20	28	27	19	17	19	16	28
11. Line Length	10	20	19	03	10	24	..	04	12	14	16	16	06	13	17	17	18	16	09	02	17
12. Judgment of Proportions	26	06	05	14	13	18	04	..	14	15	21	19	02	05	12	15	14	20	18	15	16
13. Map Distance	06	03	15	29	20	21	12	14	...	19	20	15	13	20	14	04	19	24	06	14	16
14. Path Length	15	12	14	22	15	19	14	15	19	..	18	12	18	23	21	16	15	09	12	22	10
15. Directional Orientation B	21	32	44	29	31	26	16	21	20	18	..	49	19	41	38	34	37	30	36	17	39
18. Picture Integration	32	38	36	35	28	28	16	19	15	12	49	..	11	49	38	37	41	36	37	04	44
19. Hands	11	17	14	10	12	13	06	02	13	18	19	11	..	32	32	20	15	11	13	04	08
20. Cubes	24	31	33	30	26	20	13	05	20	23	41	49	32	..	46	34	26	27	26	06	23
21. Flags, Figures, Cards	31	21	15	27	32	28	17	12	14	21	38	38	32	46	..	46	32	21	23	18	25
22. Block Counting	30	24	18	30	30	27	17	15	04	16	34	37	20	34	46	..	34	21	28	12	38
24. Pattern Analysis	27	28	25	19	24	19	18	14	19	15	37	41	15	26	32	34	..	28	31	08	36
26. Gottschaldt Figures, Part III	25	23	30	26	18	17	16	20	24	09	30	36	11	27	21	21	28	..	24	−03	32
27. Aerial Photographs, Part IV	34	31	22	18	26	19	09	18	06	12	36	37	13	26	23	28	31	24	..	08	26
29. Speed Estimation II	15	09	−06	09	22	16	02	15	14	22	17	04	04	06	18	12	08	−03	08	..	08
31. Plane Formation	23	30	27	14	10	28	17	16	16	10	39	44	08	23	25	38	36	32	26	08	..

[1] Decimal points omitted.
[2] For code numbers see table 16.30.

velocities before the testing period begins. The five velocities are then presented in random order for identification.

Plane Formation, CP805B.—This is a motion-picture test using the tachistoscopic method of presenting material to be remembered. Each presentation shows a grid of 25 squares upon which appear 5 plane silhouettes. The task of the examinee is to observe the screen and, immediately after the exposure, to fill in spaces on the answer sheet corresponding to the sections of the grid that included the planes.

Gottschaldt Figures, QP901A, Part III.—In this test the examinee's task is to determine which one of five simple geometric figures is contained in different complex geometrical figures.

Aerial Photographs, QP901A, Part IV.—This test presents a series of oblique aerial photographs upon which a number of alphabetically lettered points are placed. The examinee's task is to answer questions concerning various distance relationships between the points.

The Factors

Six factors were obtained in the first analysis. The same six factors plus two others and one residual factor were obtained in the second. In the following text, factor loadings in the two analyses are treated together. Loadings are reported if a test appears with a saturation of 0.25 or greater in either study. The rotated factors are numbered to correspond in the two analyses. The centroid loadings are shown in tables 16.29 and 16.30, and the rotated factor loadings in tables 16.31 and 16.32.

TABLE 16.29.—*Centroid factor loadings of Perceptual Battery I*[1]

Test	I	II	III	IV	V	VI	h^2
1. Speed of Identification (R), CP610A	54	−39	−25	13	−08	09	54
2. Spatial Orientation II, CP503B	51	−40	13	04	07	08	45
3. Spatial Orientation I, CP501B	55	18	−20	24	−05	12	45
4. Dial Reading, CP622A	60	29	−24	−27	11	06	59
5. Table Reading, CP621A	57	15	−36	−19	20	−14	57
6. Mathematics, B, CI206B - CI706A	46	59	−10	−18	−21	06	65
7. Mechanical Principles, CI903A	39	−09	44	−14	−22	−07	43
8. SAM Complex Coordination, CM701A	48	−22	12	−17	19	−23	41
10. Path Distance, CP608B	54	−16	−03	−11	−10	−10	35
13. Map Distance, CP626A	28	03	34	04	−13	−23	27
15. Directional Orientation B, CP515B	61	23	23	−12	12	17	53
17. Speed of Identification (N-R)	46	−39	−30	18	−10	15	52
20. Cubes, CP512A	57	04	13	−27	−13	12	45
21. Flags, Figures, Cards, CP512A	56	−17	19	−13	04	21	44
23. Path Tracing, CP512A	47	−08	−32	11	25	−12	42
24. Pattern Analysis, CP512A	56	04	04	18	15	−02	37
26. Gottschaldt Figures, QP901A, Part III	50	13	12	23	−22	−09	39
31. Plane Formation, CP805B	54	19	05	26	14	08	42

[1] Decimal points omitted.

Table 16.30.—*Centroid factor loadings of Perceptual Battery II*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	h^2
1. Speed of Identification (R), CP610A	47	08	−28	−23	17	−09	−18	06	11	44
5. Table Reading, CP621A	49	−28	09	−26	26	−20	−10	13	−08	54
6. Mathematics B, CI206B - CI706A	44	−42	30	23	−09	−15	−08	−09	05	56
7. Mechanical Principles, CI903A	44	23	−15	34	−17	−10	09	05	10	44
8. SAM Complex Coordination, CM701A	49	17	−09	−14	13	−14	06	20	−24	43
9. Point Distance, CP607B	45	11	07	−15	−18	15	11	10	05	32
11. Line Length, CP606B	29	−09	18	−17	−09	19	−04	13	14	24
12. Judgment of Proportions, CP206B	31	10	−21	−07	−24	−12	−09	−07	09	25
13. Map Distance, CP626B	34	15	10	13	−33	06	−08	18	−19	35
14. Path Length, CP628B	35	25	19	−09	−12	−11	−11	−06	09	28
15. Directional Orientation B, CP515B	66	−17	08	09	−07	−14	19	−17	−10	58
18. Picture Integration, CP104A	67	−22	−14	16	08	−08	16	13	09	61
19. Hands, CP512A	32	16	26	10	21	12	−15	−15	−03	31
20. Cubes, CP512A	60	04	17	30	28	−09	04	08	10	58
21. Flags, Figures, Cards, CP512A	60	25	05	07	25	17	13	−06	07	55
22. Block Counting, CP512A	57	08	−16	−06	15	16	23	−15	16	51
24. Pattern Analysis, CP512A	55	−15	−09	−03	02	11	−04	−11	−07	36
26. Gottschaldt Figures, OP901A, Part III	48	−18	−13	18	−15	06	−19	14	04	40
27. Aerial Photographs, OP901A, Part IV	50	−13	−18	−08	10	−14	−09	−08	−04	35
29. Speed Estimation II, CP205A II	23	31	05	−24	−10	−18	14	−09	−12	29
31. Plane Formation, CP805B	54	−24	−12	−07	−09	19	11	−04	−02	42

[1] Decimal points omitted.

TABLE 16.31.—*Rotated loadings of Perceptual Battery I*[1]

Tests	I	II	III	IV	V	VI
1. Speed of Identification (R), CP610A	19	−02	09	69	−01	−15
2. Spatial Orientation II, CP503B	40	−20	23	45	02	01
3. Spatial Orientation I, CP501B	03	28	04	51	19	24
4. Dial Reading, CP622A	41	50	−13	27	24	08
5. Table Reading, CP621A	42	54	−03	36	−04	02
6. Mathematics B, CI206C	09	60	02	10	49	14
7. Mechanical Principles, CI903A	31	−04	50	03	25	−01
8. Complex Coordination, CM701A	52	08	26	19	−14	−01
10. Path Distance, CP608B	34	17	26	35	06	−11
13. Map Distance, CP626A	14	07	47	00	07	12
15. Directional Orientation B, CP515B	47	17	12	18	36	32
17. Speed of Identification (N-R)	10	−06	03	70	−06	−14
20. Cubes, CP512A	42	17	20	21	38	−05
21. Flags, Figures, Cards, CP512A	49	−10	15	31	26	05
23. Path Tracing, CP512A	26	26	−06	48	−22	11
24. Pattern Analysis, CP512A	26	14	16	39	02	31
26. OP901A Part III (Gottschaldt figures)	−02	20	32	41	21	22
31. Plane Formation, CP805B	18	17	11	38	13	44

[1] Decimal points omitted.

TABLE 16.32.—*Rotated loadings of Perceptual Battery II*[1]

Tests	I	II	III	IV	V	VI	VII	VIII	IX
1. Speed of Identification (R), CP610A	56	16	−10	05	20	−01	15	02	11
5. Table Reading, CP621A	30	−12	05	10	35	−11	50	13	11
6. Mathematics B, CI206B - CI706A	06	15	02	13	15	−04	45	55	00
7. Mechanical Principles, CI903A	13	54	−05	02	25	04	−13	20	−05
8. Complex Coordination, CM701A	27	20	09	06	50	−03	18	−14	00
9. Point Distance, CP607B	18	21	36	23	17	05	−04	10	13
11. Line Length, CP606B	18	−01	38	11	−01	−04	09	20	08
12. Judgment of Proportions, CP206B	17	34	−05	13	07	08	00	11	23
13. Map Distance, CP626B	06	46	31	04	11	08	13	00	00
14. Path Length, CP628B	07	23	09	−11	23	17	08	21	24
15. Directional Orientation B, CP515B	02	27	−10	36	41	08	22	31	16
18. Picture Integration, CP104A	30	25	−05	28	40	−09	16	42	−12
19. Hands, CP512A	11	04	02	−05	17	46	17	09	07
20. Cubes, CP512A	24	18	−12	−03	47	25	19	41	−12
21. Flags, Figures, Cards, CP512A	33	15	05	15	43	42	−05	16	00
22. Block Counting, CP512A	42	11	−03	36	28	21	−13	20	08
24. Pattern Analysis, CP512A	25	18	−03	35	15	12	26	15	06
26. Gottschaldt Figures, QP901A, Part III	24	38	05	15	01	−02	29	27	−12
27. Aerial Photographs, QP901A, Part IV	37	18	−05	15	24	−06	27	05	12
29. Speed Estimation II, CP205A II	03	13	09	03	33	02	−08	−17	32
31. Plane Formation, CP305B	24	17	09	50	16	02	15	22	−02

[1] Decimal points omitted.

Rotated factor I is defined by the following data:

Test No.	Test name	Loadings	
		I	II
17	Speed of Identification (Nonrotated)	0.70	...
1	Speed of Identification (Rotated)	.69	0.58
3	Spatial Orientation I	.51	...
23	Path Tracing (Pursuit)	.48	...
2	Spatial Orientation II	.45	...
22	Block Counting	...	.43
26	Gottschaldt Figures, QP901A, Part III	.41	.20
24	Pattern Analysis	.39	.26
3	Plane Formation	.38	.23
27	Aerial Photographs, QP901A, Part IV	...	.39
5	Table Reading	.36	.30
10	Path Distance	.25	...
21	Flags, Figures, Cards	.31	.31
18	Picture Integration	...	.27
8	Complex Coordination	.29	.25
4	Dial Reading	.27	...

This factor is clearly the perceptual-speed factor, usually defined by the Speed of Identification test. In the first analysis, probably because of the presence of two forms of the same test, the loading for Speed of Identification is higher than usual. In the second analysis, the loading is somewhat below the mean. It is possible that in the first analysis there is specific nonerror variance of the test included in the factor loading. One noteworthy discrepancy between the loadings in these two analyses is for Plane Formation. In the second analysis the rotations reduced the loading on this factor and increased the loading on the visual-memory factor in which it is more heavily saturated.

Rotated factor II is defined by the following data:

Test No.	Test name	Loadings	
		I	II
7	Mechanical Principles	0.50	0.54
13	Map Distance	.47	.42
26	Gottschaldt Figures	.32	.38
12	Judgment of Proportions	...	.32
10	Path Distance	.26	...
18	Picture Integration	...	.26
8	Complex Coordination	.26	.19
15	Directional Orientation B	.12	.26

Because this factor is best defined by tests that require the manipulation of visual images, it has been termed visualization. Other tests that measure the factor are Pattern Comprehension and Spatial Visualization I and II.

Rotated factor III is defined by the following data:

Test No.	Test name	Loadings	
		I	II
31	Plane Formation	0.44	0.50
15	Directional Orientation	.32	.36
22	Block Counting	...	.36
24	Pattern Analysis	.31	.35
18	Picture Integration	...	.28

This factor is not well known. It has been suggested that pattern perception might be the fundamental element involved. It is possible, however, that a memorial component is the single feature that these tests have in common. Similarities to new factors obtained in other analyses give some support for this conclusion. A clear-cut visual-memory factor was isolated in the analyses of the Memory Batteries (see ch. 11). The Map Memory tests have substantial loadings on the factor. Although further work will be necessary before these factors can be identified with each other with complete assurance, the factor will tentatively be labeled "visual memory."

Rotated factor IV is described by the following data:

Test No.	Test name	Loadings	
		I	II
6	Mathematics B	0.60	0.45
5	Table Reading	.54	.50
4	Dial Reading	.50	...
26	Gottschaldt Figures	.20	.29
3	Spatial Orientation I	.28	...
27	Aerial Photographs	...	.27
25	Path Tracing (Pursuit)	.26	...
24	Pattern Analysis	.14	.26

This is the familiar numerical factor, best defined by the Numerical Operations test. The absence of other number tests from the correlational matrix made the differentiation between number and reasoning difficult. The loadings of the two classification tests (Mathematics B, and Dial and Table Reading) are congruent with previous results, but it is difficult to rationalize the correlations of the three experimental tests with this factor in the second analysis. In the first study the experimental tests are less numerical but have higher reasoning loadings.

Rotated factor V is defined by the following data:

Test No.	Test name	Loadings	
		I	II
6	Mathematics B	0.49	0.55
18	Picture Integration	...	.42
20	Cubes	.38	.41
15	Directional Orientation	.36	.31
26	Gottschaldt Figures	.21	.27
21	Flags, Figures, Cards	.26	.16
7	Mechanical Principles	.25	.20
4	Dial Reading	.24	...

This is the general-reasoning (reasoning I) factor. It was not clear at the time the factor was first isolated whether or not the modifying term "inductive" should be used in connection with its description. Many tests, clearly inductive in character, do have high loadings on the factor. It is also probable that at least two other reasoning factors have been isolated (see ch. 7).

Rotated factor VI is defined by the following data:

Test No.	Test name	Loadings I	Loadings II
8	Complex Coordination	0.52	0.50
21	Flags, Figures, Cards	.49	.43
20	Cubes	.42	.47
15	Directional Orientation	.47	.41
5	Table Reading	.42	.35
4	Dial Reading	.41	...
18	Picture Integration	...	.40
2	Spatial Orientation II	.40	...
10	Path Distance	.34	...
29	Speed Estimation II	...	.33
7	Mechanical Principles	.31	.25
22	Block Counting	...	.28
23	Path Tracing	.26	...
24	Pattern Analysis	.26	.15

This factor has been termed spatial relations. It was surprising to find, in the second analysis, that Thurstone's Hands test appeared projected on a different factor (factor VII). Apparently two spatial factors are involved; one is best defined by Complex Coordination, the other by the Hands test. Flags and Cubes appear on both. Other tests known to be loaded with the original spatial-relations factor are Instrument Comprehension II, Two-Hand Coordination, and Discrimination Reaction Time.

Rotated factor VII is defined by the following data from the second analysis only:

Test No.	Test name	Loading II
19	Hands	0.46
21	Flags, etc.	.42
20	Cubes	.25

This is the factor just discussed. Since a rotation or a positional change seems to be involved, the factor has been tentatively described as "rotational" or "positional space." Another hypothesis is that the common element enabling the subjects to solve problems readily is the ability to enter the self into the action; that is, by empathetic involvement, which would call for the name "spatial empathy." In order not to prejudge the factor, however, it is designated as space II.

Rotated factor VIII is defined by the following tests from the second analysis only:

Test No.	Test name	Loadings II
11	Line Length	0.34
9	Point Distance	.39
13	Map Distance	.31
29	Speed Estimation II	.29
14	Path Length	.25

This factor has a precedent in the one called length estimation found in the analysis of the Mechanical Battery (see ch. 13). The factor is common to the Path Distance test and the Pattern Assembly test, as well as those appearing here.

Rotated factor IX is a residual factor.

Conclusions

The most interesting areas for further test construction are the ones defined by rotated factor III (visual memory) and by rotated factor VII (Space II). Discussion of the visual-memory factor may be found in chapter 11 in the report of the factor analysis of memory tests and in chapter 12 in the evaluation of visualization tests. Space II is mentioned in chapter 19 in the discussion of evaluation of space tests.

It is of general interest to find that out of all the variety of perceptual tests analyzed, only two clearly perceptual factors—perceptual speed and length estimation—emerged. A possible third is a pattern-perception factor, but that hypothesis was rejected in favor of calling it visual memory. None of the new factors reported by Thurstone (3) was brought out. In a number of the tests, the nonerror variances are not fully accounted for, however, so that in more favorable batteries some of Thurstone's new factors may yet come to light in these tests.

BIBLIOGRAPHY

(1) Thurstone, L. L. *Primary Mental Abilities*, Chicago, University of Chicago Press, 1938.

(2) Thurstone, L. L. The perceptual factor. *Psychometrika*, 1938.

(3) Thurstone, L. L. *A Factorial Study of Perception*, Chicago, University of Chicago Press, 1944.

CHAPTER SEVENTEEN

Form Perception Tests[1]

INTRODUCTION

In this chapter are discussed tests of form perception under the following a priori rubrics: Pattern-formation tests, pattern-completion tests, pattern-analysis tests, and illusions tests.

In the pattern formation tests, the examinee is required to reorganize disordered segments into a coherent whole. The pattern-completion tests present mutilated patterns which the examinee must recognize. In the pattern-analysis tests, faces, letters, nonsense symbols, and words are concealed in a complex, camouflaging context. The examinee must form correct figure-ground relationships. The illusions tests present examinees with familiar geometric and Gestalt illusions and attempt to measure susceptibility to these illusions.

With some exceptions, these tests were not constructed in the light of specific job-analysis information, but rather because of systematic interests and psychologically guided speculation.

PATTERN FORMATION TESTS

Picture Integration, CP104A[2]

This test was designed as a measure of the ability to visualize objects in space and as a test of perceptual integration, on the assumptions that visualization of spatial relationships and the ability to perceive a scene adequately from fragmentary or distorted cues are important for pilots.

Description.—Each test item consists of a photograph which has been cut into four quarters and rearranged in a scrambled order. One of the sample items used in the directions is shown in figure 17.1. The task of the examinee is to visualize the correct order of the disarranged segments and to indicate the correct arrangement on a work sheet. In the bottom panel of figure 17.1 is shown an answer box correctly completed for the illustrated item.

(1) *Internal characteristics.*—The directions include four sample problems. There are 30 test items, arranged 6 to a page. The total number of scored responses, then, is 120.

(2) *Administration.*—The directions to the test require approximately 5 minutes. The test items originally required another 20 minutes, but it

[1] Written by Capt. John I. Lacey. Sgt. Stanley W. Niehaus and Tech/Sgt. Gerald H. Shirley assisted in collating the materials for this chapter.

[2] Developed at the Perceptual Research Unit, Headquarters, AAF Training Command. Chief contributors: Capt. Richard H. Henneman and staff.

was found that 13 minutes suffice. Upon completion of the test, the examinees transcribe their answers on the work sheet to a specially prepared IBM answer sheet.

(3) *Scoring.*—The scoring formula is (3R−W)/4.

Statistical results.—The data given below are mainly for examinees tested in April 1942 at Psychological Research Unit No. 3; those who went to pilot training were in class 43K. Exceptions are noted in the appropriate places.

(1) *Distribution statistics.*—A sample of 392 unclassified aviation students yielded a mean score of 53.5 and a standard deviation of 17.3.

(2) *Internal consistency.*—The degree of homogeneity of the *responses* (there are four responses to an item) is indicated by a mean phi coefficient of 0.32, with a range from 0.07 to 0.56, and a standard deviation of 0.08. These statistics are based upon the responses of the highest 27 percent and the lowest 27 percent of a group of 350 unclassified aviation students.

(3) *Reliability coefficient.*—The reliability was estimated by correlating the score on the first 2 pages (first 48 responses) with the score on the last 3 pages (last 72 responses). For experimental purposes, these artificial parts were timed separately, with an allowance of 4½ and 8 minutes, respectively. The correlation was 0.63 for 422 pilots. A very rough estimate of the reliability of the total test, then, is 0.77.

(4) *Difficulty.*—Based upon the responses of 350 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.76, corrected for chance success, with a range from 0.47 to 0.92 and a standard deviation of 0.12.

(5) *Factorial composition.*—The most prominent loadings were found in the general-reasoning (0.42), spatial-relations (0.40), visual-memory (0.28), perceptual (0.27), and visualization (0.26) factors. The communality was 0.61. For a fuller picture of the factorial composition of the test, see Appendix B.

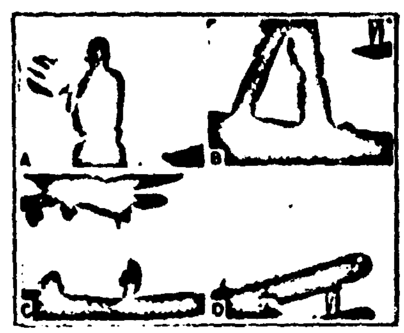
(6) *Test validity.*—Validation results based on overlapping samples are presented in table 17.1.

TABLE 17.1.—*Validation data for Picture Integration, CP104A, for pilots in primary training, graduation-elimination criterion*

N_t	p_s	M_s	M_u	SD_t	r_{bis}	$_c r_{bis}$ [a]
[b] 302	0.87	54.35	50.45	15.25	0.17	0.21
[c] 341	.87	54.30	49.10	15.15	.18	...
[d] 359	.84	56.03	50.65	14.66	.20	.28

[a] Assuming an unrestricted stanine standard deviation of 2.00.
[b] In class 44B.
[c] In classes 44B and 44C. Includes the first sample.
[d] In classes 43K, 44A, 44B, and 44C. Overlaps with first two samples.

Evaluation.—This test has a satisfactory reliability and moderate validity for pilots.



A	D
B	C

FIGURE 17.1

SAMPLE ITEM & ANSWER BOX OF PICTURE INTEGRATION.- CP104A

The pilot validity estimated from factorial equations (see ch. 28) is 0.22, which is close to the mean obtained validity of 0.25. A moderate navigator validity probably would be found for this test also, because of its saturations in the general-reasoning, spatial-relations, and perceptual-speed factors.

Sixty-one percent of the total variance of the test has been accounted for by identified factors. This probably leaves a considerable amount of undefined nonerror variance, which may be found to be attributable to some new perceptual factor. The usefulness of the test, however, in factor research, or in a classification and selection program that utilizes pure tests, is limited because of its factorial complexity. The following factors account for a substantial part of the total test variance: General reasoning, 18 percent; spatial relations, 16 percent; visual memory, 8 percent; perceptual-speed, 7 percent; and visualization, 7 percent. The intention of the test constructors to measure visualization, therefore, was not accomplished; the test's maximum loading is in the general-reasoning factor.

Pattern Assembly, CP804A [8]

This test was originally developed for inclusion in an experimental battery of mechanical-comprehension tests. It is a variation of the familiar paper form-board test, which has seen much use in industrial psychology as a component of mechanical-ability test batteries. Its correlations with other mechanical-comprehension tests in the experimental mechanical battery, however, indicated that it did not have much in common with them. The test, therefore, was assigned a perceptual code number. Like the Picture Integration test, it seemed to require visualization and the integration of the disordered parts of a whole.

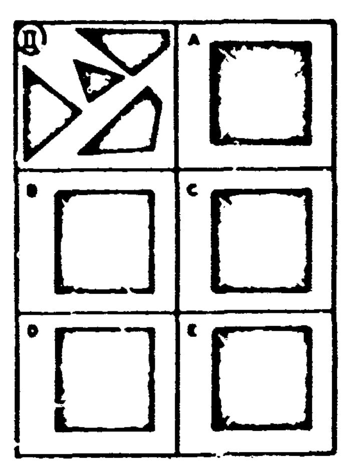


FIGURE 17.2
SAMPLE ITEM OF PATTERN ASSEMBLY, CP804A

[8] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. Lewis G. Carpenter Jr., and Tech/Sgt. Paul C. Davis.

703120—47—28

Description.—The nature of the test items is best explained by reference to figure 17.2, which is one of the sample items used in the directions to the test. The examinee is required to select that one of the five patterns lettered A to E which shows exactly how the parts shown in the upper left-hand corner would look when fitted together. The correct answer is B. It is important to note that one of the four pieces must be turned over in order to fit into pattern B.

(1) *Internal characteristics.*—The test includes 34 items, the first 2 of which are unscored sample items.

(2) *Administration.*—Five minutes are allotted to administration of directions and 15 minutes for answering the test items.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data given are for examinees tested in November 1942 at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—A sample of 485 unclassified aviation students yielded a mean score of 15.1 and a standard deviation of 5.1.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean phi coefficient of 0.26, with a range from 0.10 to 0.54 and a standard deviation of 0.09. These data are based upon the responses of the upper 50 percent and the lower 50 percent of a group of 150 unclassified aviation students.

(3) *Reliability coefficient.*—By the alternate-forms method (artificially separated halves, separately timed), an estimated reliability coefficient of 0.59, corrected for length, was obtained. This figure is based on a sample of 202 unclassified aviation students.

(4) *Difficulty.*—Based upon the responses of 150 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.64, corrected for chance success, with a range from 0.12 to 0.94 and a standard deviation of 0.26.

(5) *Factorial composition.*—This test was included in only one factor analysis. Prominent loadings were found only in the length-estimation (0.52) and perceptual-speed (0.31) factors. Slight loadings were found also in the visualization (0.16) and general-reasoning (0.14) factors. The communality is 0.42. For a fuller picture of the factorial composition of this test, see Appendix B.

(6) *Test validity.*—Validity data are presented in table 17.2.

TABLE 17.2.—*Validity data for Pattern Assembly, CP501A and CP504AX1, for pilots in primary training, graduation-elimination criterion*

Form	N	Score	p_t	M_p	M_e	SD_t	r_{bis}
CP501A ...	[1]533	R−W/4 ..	0.92	16.62	13.90	5.08	0.25
CP501A ...	[2]607	R−W/4 ..	.81	15.98	14.50	5.20	.22
CP504AX1[3]	154	R	.74	22.24	21.65	4.70	.07
CP504AX1[3]	154	R−2W ...	.74	8.93	6.92	8.79	.14

[1] In class 43I.
[2] In class 43K.
[3] A very slightly different form; see below.

A variation: CP804AX1.—This, the original form of the Pattern Assembly test, is composed of 40 items, most of which are identical with those of the revised form, CP804A.

Evaluation.—Of all tests studied in the program, this test has the highest loading (0.52, accounting for 27 percent of the total test variance) in the length-estimation factor. It, therefore, deserves serious consideration. Its next highest loading (0.31, accounting for 10 percent of the test's variance) is in the perceptual-speed factor, and, probably, this can be decreased by making the details of the test items more easily discriminable as to form and detail. The very low loadings on the visualization (0.16) and general-reasoning (0.14) factors are encouraging (to those interested in pure tests). These loadings, indeed, may represent sampling deviations from zero. If they do not, the visualization loading, at least, probably can be decreased by not requiring that some of the segments be turned over before they can be fitted into the key figure. Since only 42 percent of the total test variance is accounted for by the factors so far identified in this test, however, there is considerable non-error variance yet undefined. There is room for a new factor with a loading in this test of approximately 0.40.

The pilot validity is only moderate, the weighted average validity for a total of 839 cases being only 0.18. The validity predicted from factorial equations (see ch. 28) is 0.16, which agrees well with the obtained validity.

In its present form, the test is slightly too easy, and its reliability is only minimally satisfactory, although it could be a useful member of a battery of tests.

Area Visualization, CP815A [a]

This test was designed as a measure of the ability of manipulatory visualization. It is somewhat similar to Pattern Assembly, CP804A.

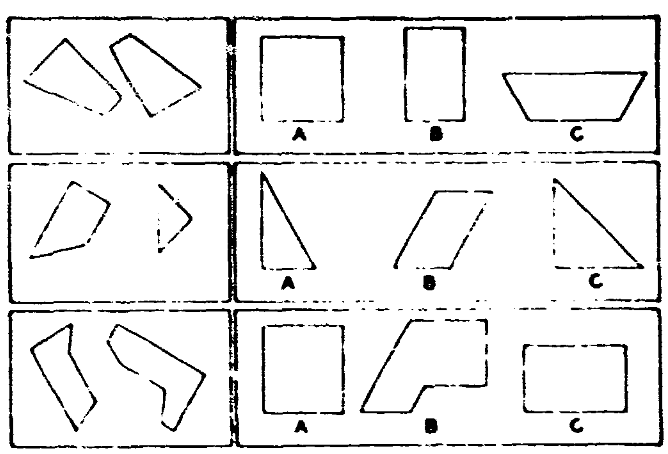


FIGURE 17.3
SAMPLE ITEMS OF AREA VISUALIZATION. CP815A

[a] Developed at Psychological Research Unit No. 1. Chief contributor: Pfc. T. Blake.

Description.—For each test item, the examinee is required to indicate which one of three figures will be formed when two segments are rotated about so that they fit together. Sample items are shown in figure 17.3.

(1) *Internal characteristics.*—The test is divided into two parts of 30 items each. One other item is used as an illustrative sample in the directions. Unlike the Pattern Assembly test, segments need not be turned over; all that is required is rotation.

(2) *Administration.*—Each part is allotted 7 minutes, and the directions take another 5 minutes.

(3) *Scoring.*—The scoring formula is $R-W/2$.

Evaluation.—Since there are no statistical data available for this test, little can be said in evaluation of it. The test is relatively long, and it should have satisfactory reliability. Inspection of the test items suggests that length-estimation variance is reduced sharply, as compared with Pattern Assembly, CP804A.

It seems doubtful that the test will have much visualization variance, judging from its similarities to the Pattern Assembly test.

PATTERN COMPLETION TESTS

Mutilated Words, CP512A [a]

This is one of a battery of nine tests adapted from forms devised by L. L. Thurstone. These tests were intended to measure visualization abilities.

Description.—Each item consists of a mutilated word, i. e., a printed word partially erased, and five complete words labeled A through E. It is the task of the examinee to recognize the incomplete word and to select from five alternatives the word that bears the closest relationship to it. The completed form of the mutilated word is not presented among the five alternatives, but the answer word is sufficiently similar in meaning to be easily recognized.

Two sample test items are shown in figure 17.4.

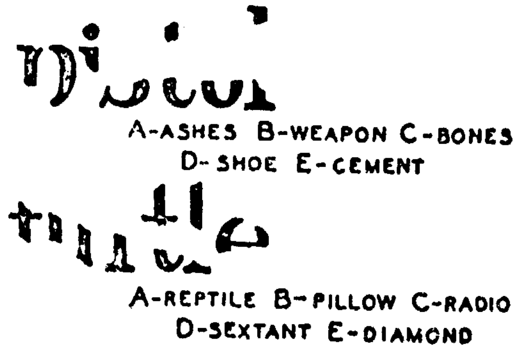


FIGURE 17.4
SAMPLE ITEMS OF MUTILATED WORDS, CP512A

[a] Developed at Perceptual Research Unit, Headquarters, AAF Training Command. Chief contributors: Capt. Richard H. Henneman and staff.

(1) *Internal characteristics.*—The test is made up of 5 sample and 25 scored items. They are arranged in approximate order of difficulty.

(2) *Administration.*—The over-all testing time is approximately 9 minutes, with 4 minutes required for the test proper and 4 to 5 minutes for directions and sample problems.

(3) *Scoring.*—The score on this test is simply the number of correct responses.

Statistical results. (1) *Distribution statistics.*—Distribution statistics obtained on this test are given in table 17.3. The distribution curves are moderately negatively skewed.

TABLE 17.3.—*Distribution constants for Mutilated Words, CP512A*

Group	N	M	SD
Unclassified aviation students[1]	460	17.8	3.1
Classified pilots[2]	640	17.8	3.1
Classified pilots[3]	185	18.2	2.7

[1] Tested in March 1943 at Psychological Research Unit No. 3.
[2] Tested in April 1943 at Psychological Research Unit No. 1. In classes 44A, 44B, and 44C.
[3] Tested in August 1943 at Psychological Research Unit No. 3. In class 43K.

(2) *Test validity.*—Validation results based on several samples are given in table 17.4.

TABLE 17.4.—*Validity data for Mutilated Words, CP512A, based upon samples of pilots in primary training, graduation-elimination criterion*

N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
[2]640	0.74	17.92	17.64	3.11	0.05	0.08
[3]185	.86	18.21	18.12	2.71	.00	...

[1] Assumed unrestricted stanine standard deviation not reported.
[2] In classes 44A, 44B, and 44C. Tested in April 1943 at Psychological Research Unit No. 1.
[3] In class 43K. Tested in August 1943 at Psychological Research Unit No. 3.

Evaluation.—The paucity of data for this test makes evaluation difficult. Because of the small number of items, it probably is not very reliable, and the distribution curves show that the items are relatively easy. The test has almost no pilot validity, so we may be rather sure that none of the factors known to be valid for pilots is substantially represented. In Thurstone's analysis (2) of a 15-item form of the test (in which the examinee responded directly by reading the mutilated word), the principal loadings of the test were in a factor identified as "speed and strength of closure," and in another factor called "speed of perception." The latter, best defined in Thurstone's analysis by peripheral span and dark-adaptation measures, probably is not the same as the perceptual-speed factor defined in the present volume.

Object Completion, CP811A[4]

This test was designed specifically to measure the ability to perceive the form of an object when only a portion of its elements can be seen.

[4] Developed at Psychological Research Unit No. 3. Chief Contributors: Tech/Sgt. Paul C. Davis, Capt. Richard H. Henneman, Sgt. Frederick H. Meise, and Tech/Sgt. Sanford J. Mock.

It was thought that perceptual integration, or the ability to recognize total situations from partial impressions, as in the identification of terrain from fleeting, incomplete glimpses through clouds or smoke, might be an important function in air-crew success.

Description.—The test is a modification of the Street Gestalt completion test (1). It consists of a series of drawings of military objects from which many of the parts are deleted. It is the examinee's task to select from a list of alternative answers the name of the object which is partially portrayed in the item.

(1) *Internal characteristics.*—On each double-page spread, six incomplete drawings are presented. With these 6 pictures is a list of 13 possible answers, including the correct ones, misleads, and other. Figure 17.5 shows some items of the test. The test is divided into 2 parts of 30 items each. In addition, there is one sample item used in the directions.

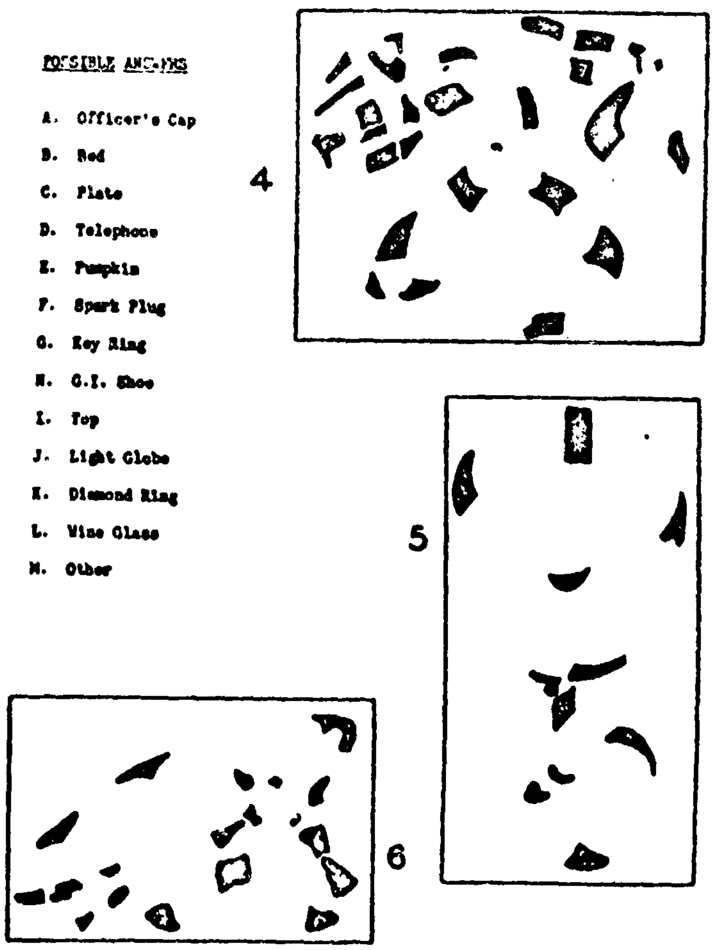


FIGURE 17.5
SAMPLE ITEMS OF OBJECT COMPLETION, CP811A

(2) *Administration.*—The sample item is explained and answered before the test is begun. Fifteen minutes are allowed for each part. Administration requires approximately 5 minutes.

(3) *Scoring.*—The scoring formula is R—W/4.

Statistical results. (1) *Distribution statistics.*—In table 17.5 are shown distribution data for unclassified aviation students and for classified pilots. The distribution curves are slightly positively skewed.

TABLE 17.5.—*Distribution constants for Object Completion, CP811A*

Group	Part	Score	N	M	SD
Classified pilots[1]	I	R	563	13.6	6.1
Do[1]	I	R	1,310	12.7	5.5
Do[1]	II	R	563	12.3	6.3
Do[1]	II	R	1,310	14.3	5.3
Unclassified aviation students[2]	I and II	R	500	27.4	9.6
Classified pilots[2]	I and II	R	339	26.4	9.4
Do[1]	I	W	1,310	10.3	5.7
Do[1]	II	W	1,310	8.6	5.1
Do[2]	I and II	W	339	18.1	10.4
Unclassified aviation students[2]	I and II	W	500	17.1	10.0

[1] In class 44J. Tested in April 1944 at Psychological Research Unit No. 3.
[2] Tested at Psychological Research Unit No. 3 in May 1944.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.37, with a standard deviation of 0.14 and a range of values from 0.10 to 0.66. These statistics are based upon the upper 27 percent and the lower 27 percent of a group of 570 unclassified aviation students, tested in May 1944 at Psychological Research Unit No. 3.

(3) *Reliability coefficients.*—Samples to which this test was administered yielded the estimates of reliability given in table 17.6.

TABLE 17.6.—*Estimated reliability coefficients for Object Completion, CP811A, based upon samples of classified pilots*[1]

N	Variables	r'_{tt}	r_{tt}
1,378	Part I (R) v. Part II (R)[2]	0.68	0.81
563	Part I (R) v. Part II (R)	.75	.86

[1] In class 43J. Tested at Psychological Research Unit No. 3 in April and May 1944. Overlapping samples.
[2] (R) indicates a simple "number right" score.

(4) *Correlation between rights and wrongs.*—Based upon a sample of 500 unclassified aviation students tested at Psychological Research Unit No. 3 in May 1944, the correlation between total rights and total wrongs was −0.37.

(5) *Difficulty.*—Based upon the responses of 750 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.56, with a standard deviation of 0.16 and a range from 0.05 to 0.81. This sample was tested in May 1944 at Psychological Research Unit No. 3. Correction for chance success was not attempted, because the test is a matching test with progressive elimination of the number of alternative responses.

(6) *Test validity.*—Validation results based on several samples are given in table 17.7.

TABLE 17.7.—*Validity data for Object Completion, CP811A, based upon samples of pilots in primary training, graduation-elimination criterion*

Score	N_t	r_t	M_p	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
Part I (R) ..	[2]339	0.76	13.04	11.28	5.2	0.20	0.26
Part I (R) ..	[3]1,310	.90	12.73	12.76	5.46	−.003	.07
Part I (W) .	[2]339	.76	9.24	10.19	5.7	−.10	−.11
Part I (W) .	[3]1,310	.90	10.34	10.33	5.73	.02	−.03
Part II (R)	[2]339	.76	14.07	13.08	5.1	.11	.16
Part II (R)	[3]1,310	.90	14.31	14.00	5.29	.03	.10
Part II (W)	[2]339	.76	8.52	8.96	5.4	−.05	−.05
Part II (W)	[3]1,310	.90	8.59	8.91	5.12	−.03	−.06
Total (R)	[2]339	.76	27.11	24.36	9.3	.17	.23
Total (W) ...	[2]339	.76	17.76	19.15	10.3	−.08	−.09

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Tested May 11 to May 15, 1944, at Psychological Research Unit No. 3.
[3] Tested in April 1944 at Psychological Research Unit No. 3. In class 44J.

Evaluation.—This test does not have much promise as a pilot-selection instrument. Its reliability and difficulty level are satisfactory. In Thurstone's analysis (2) of perceptual tests, another modification of the Street Gestalt completion test has prominent loadings in two perceptual factors: speed and strength of closure, and speed of perception.

PATTERN ANALYSIS TESTS

Pattern Analysis, CP512A [7]

This is a variant of the Gottschaldt concealed-figure test utilizing only one standard figure (a capital Greek sigma) embedded in various complex designs. The test was designed to measure the ability to form figure-ground relationships, and, like Mutilated Words discussed above, is one of the battery of nine tests designed for the analysis of visualizing abilities.

Description.—The task of the examinee is to detect the outline of the standard figure in a complex design. Figure 17.6 shows the standard figure and five sample items. The standard figure can be detected in alternatives A, C, and D. (In the actual test, of course, the standard figure is not shown with every set of items. It is shown only once, in the directions to the test.)

(1) *Internal characteristics.*—Test instructions include a showing of the standard figure, 18 exemplary items, and 15 practice items. The test proper consists of 1,035 items, arranged in blocks of five. If the design contains the standard figure, the examinee is required to fill in the appropriate answer-space; if it does not, the examinee leaves the answer-space blank. The standard figure actually appears in 104 of the items.

(2) *Administration.*—Administration requires approximately 5 minutes, with an over-all testing time of 17 minutes. In the original admin-

[7] Developed at Perceptual Research Unit, Headquarters, AAF Training Command. Chief Contributors: Capt. Richard H. Henneman and staff.

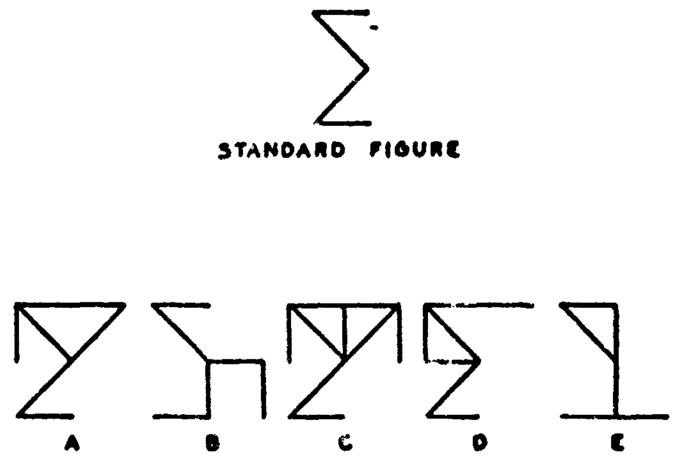


FIGURE 17.6
STANDARD FIGURE & SAMPLE ITEMS OF PATTERN ANALYSIS, CP512A

istration of these tests, however, 15 minutes were allowed for the first 735 items.

(3) *Scoring.*—The scoring formula is R−W.

Statistical results. (1) *Distribution statistics.*—In table 17.8 are presented distribution constants for two samples of classified pilots.

TABLE 17.8.—*Distribution constants for Pattern Analysis, CP512A, based upon samples of pilots*

N	M	SD
[1]392	59.3	18.7
[2]640	64.4	16.6

[1] Tested at Psychological Research Unit No. 3. Class not reported.
[2] Tested at Psychological Research Unit No. 1 in April 1943 in classes 44A, 44B, and 44C.

(2) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.91, corrected for length, was obtained. This figure is based on a sample of 438 unclassified aviation students and on the administration of separately timed halves of the test.

(3) *Factorial composition.*—The most significant loadings are in the visual-memory (0.35), perceptual-speed (0.26), numerical (0.26), and visualization (0.18) factors. The communality is 0.35. For these factor analyses, only the first 735 items of the test were administered. (The estimated reliability is 0.87, based on 460 unclassified aviation students and on administration of separately timed halves.) For a fuller picture of the factorial composition of this test, see Appendix B.

(4) *Test validity.*—A sample of 640 pilots (classes 44A, B and C; see table 17.8) yielded a biserial correlation of 0.16, corrected for re-

striction of range, between performance in this test and the graduation-elimination criterion from primary training. The mean score for graduates was 65.30, for eliminees 62.00; and the standard deviation for both combined was 16.61. Of this sample, 74 percent were graduated, and the standard deviation assumed for the unrestricted pilot stanine distribution was 2.00

Evaluation.—This test is highly reliable, but its validity for pilots is low. Only 35 percent of the total test variance is accounted for by the factors so far identified in the test, which leaves a very large amount of undefined nonerror variance.

The identified factors account for the following percentages of the total variance of the test: Visual memory, 12 percent; perceptual speed, 7 percent; numerical, 7 percent; and visualization, 3 percent. No other factor accounts for more than 2 percent of the test's variance.

Based on factorial equations (see ch. 28), the predicted pilot validity is 0.15, which indicates that all the test's pilot validity is accounted for.

Camouflaged Outlines, CP821A [8]

This is the final variation of the Gottschaldt figures test developed in the AAF program. An earlier form had been developed before the war by Col. J. P. Guilford. It was designed for inclusion in a special experimental study, which was interrupted by the end of the war. While no statistical data are available, its description is of some interest, since experience with previous experimental forms of this type of test determined its specific characteristics.

An original nonmachine-scorable form of the Gottschaldt figures test was administered as Hidden Figures, CP802A. For a group of 652 pilots, this test correlated 0.36 with graduation-elimination through advanced training. Following this lead, extensive investigations were made into the usefulness of this type of test by those responsible for the development of the AAF Qualifying Examination.[9]

The results indicated (*a*) that very easy items should be avoided, (*b*) that the test had low to moderate pilot validity (ranging from 0.17 to 0.25 in different samples), (*c*) that satisfactory results were attained with either closely timed or essentially untimed administration, and (*d*) that an extensive practice period was necessary.

In developing Camouflaged Outlines, CP821A, all previously constructed test items were carefully scrutinized. Many were used in their original form, others were modified, and some were constructed especially for the purposes of this test. An attempt was made to have all items close to the 50 percent difficulty level, using a priori judgments, available item statistics, and pretesting with members of the test-construction staff.

Description.—As in other tests of this type, the examinee is required

[8] Developed at Psychological Research Unit No. 2. Chief contributors: Capt. John I. Lacey, Jeanette B. Russell.

[9] These results are not presented in this report. For a detailed discussion, see Report No. 6 in this series, *The AAF Qualifying Examination.*

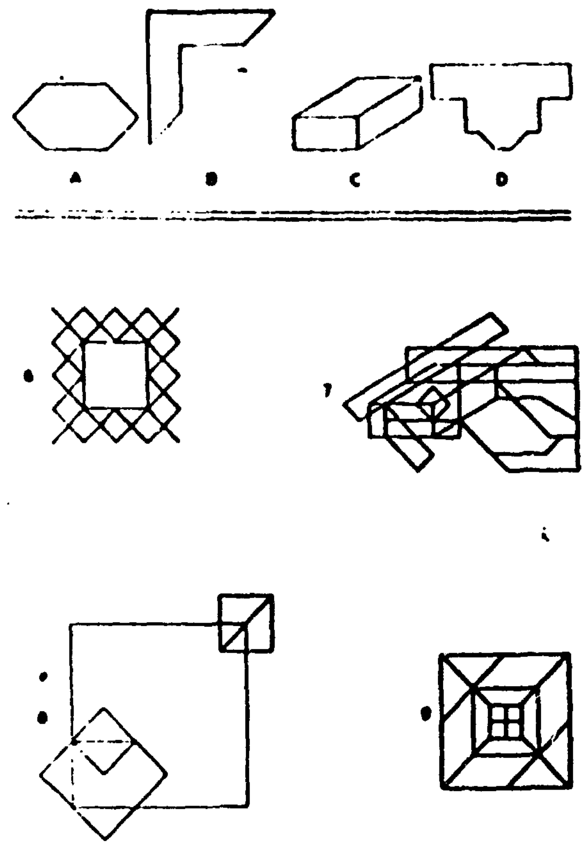


FIGURE 17.7
STANDARD FIGURES & TEST ITEMS OF CAMOUFLAGED OUTLINES, - CP821A

to detect a simple figure within a complex design. Two sets of four standard figures each are provided. Figure 17.7 shows in the top panel one set of standard figures, and, in the lower panel, four test items. The examinee is required to indicate which of the standard figures is in cluded in each complex design.

(1) *Internal characteristics.*—Extensive directions, including prac tice items, are utilized. The examinee is shown one item with a heavy outline showing a standard figure. He then attempts four practice items. After these four items are answered, the correct answers are given to him, again utilizing heavy outlines of the standard figures. The test proper is divided into 2 parts of 16 items each. In the first part, the first eight items involve the first set of standard figures, the second eight items, the second set of standard figures. In part II the order of presentation of the standard figures is reversed, thus providing an ABBA order.

(2) *Administration.*—Five and one-half minutes are allowed for each part, with approximately an additional 5 minutes required for administration of the directions.

(3) *Scoring.*—The scoring formula is R−W/3.

Penetration of Camouflage, CP812A[16]

This test was adapted from Thurstone's Hidden Pictures test.

Description.—The test consists of six page-size drawings depicting military activities. Concealed faces are worked into the context of each scene. The faces may be front view, profile, right side up, or upside down. If necessary to detect the faces, the examinees may turn the booklet in any direction. The left and right borders of the pictures are blocked off into five sections, which correspond to five numbered item spaces on the IBM answer sheet. The upper and lower borders are blocked off into five equal sections which correspond to the five lettered alternatives for each item-number. The task of the examinee is to detect the camouflaged faces in each picture and to mark their location in terms of the item-alternative coordinates. Figure 17.8 shows the sample scene used in the administrative directions. In this sample, some of the concealed figures are encircled.

FIGURE 17.8
SAMPLE DRAWING OF PENETRATION OF CAMOUFLAGE, CP812A

(1) *Administration.*—The examinees are told that: "This is a test of your ability to detect camouflaged figures * * * DO NOT indicate on

[16] Developed at Psychological Research Unit No. 3. Chief contributors: Sgt. Frederick H. Meise, Lois G. Wright.

your answer sheet the obvious faces of people in the pictures. * * *" The faces in the first two item sections of the sample scene are encircled, and the examinees are told to detect those in the last three sections. After this is done, the locations of the faces are pointed out. Two and one-half minutes are allowed for each picture. At the end of that time the examinees are told to start on the next. The approximate total testing time is 25 minutes. The first 3 pictures are considered to constitute part I, and the last three, part II, with 36 and 35 concealed faces respectively.

(2) *Scoring.*—Rights and wrongs are scored separately for each part.

Statistical results. (1) *Distribution statistics.*—Distribution statistics obtained for this test are given in table 17.9.

TABLE 17.9.—*Distribution constants for Penetration of Camouflage, CP812A, based upon a sample of 773 pilots*[1]

Score	M	SD
Rights	43.4	8.5
Wrongs	2.2	2.9

[1] Tested June 12 to Aug. 11, 1944, at Psychological Research Unit No. 3.

(2) *Internal consistency.*—The degree of homogeneity of the items of the test is indicated by a mean internal-consistency phi of 0.32 and a range of values from 0.00 to 0.58. These statistics are based upon the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 740 unclassified aviation students, tested in June 1944 at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—A sample to which this test was administered yielded the estimates of reliability given in table 17.10.

TABLE 17.10.—*Estimated reliability coefficients for Penetration of Camouflage, CP812A, based upon a sample of 773 pilots*[1] *in primary training*

Variables	r'_{tt}	r_{tt}
Part I rights v. Part II rights	0.68	0.81
Part I wrongs v. Part II wrongs	.55	.71

[1] Tested June 12 to Aug. 11, 1944, at Psychological Research Unit No. 3.

(4) *Correlations of rights and wrongs.*—The intercorrelations of rights and wrongs are shown in table 17.11.

TABLE 17.11.—*Intercorrelations of rights and wrongs of Penetration of Camouflage, CP812A, for a sample of 773 pilots in primary training*[1]

	1	2	3	4
1. Part I rights	...	0.68	−0.02	0.05
2. Part II rights	0.68	...	.06	.11
3. Part I wrongs	−.02	.06	...	.55
4. Part II wrongs	.05	.11	.55	...

[1] See footnote 1, table 17.10.

(5) *Difficulty.*—Based upon item analysis of the responses of 750 unclassified aviation students, tested in June 1944 at Psychological Research Unit No. 3, the test yielded a mean proportion of correct responses of 0.68, with a range from 0.15 to 0.99 and a standard deviation of 0.21.

(6) *Test validity.*—Validation results based on a sample of 773 pilots in elementary training are given in table 17.12.

TABLE 17.12.—*Validity data for Penetration of Camouflage, CP812.1, for pilots in primary training, graduation-elimination criterion (N_t=773[1])*

Part	Score	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[2]
I	Rights	22.96	23.12	5.13	−0.02	0.02
II	....do	20.48	20.21	4.17	.04	.10
I	Wrongs	.74	.99	1.19	−.13	−.15
II	....do	1.37	1.43	2.13	−.02	−.04
I and II	Rights	43.44	43.33	8.53	.01	.06
I and II	Wrongs	2.11	2.42	2.95	−.06	−.08

[1] See footnote 1, table 17.10.
[2] Assuming an unrestricted stanine standard deviation of 2.00.

Evaluation.—Both rights and wrongs for this test have satisfactory reliability, and they apparently measure different functions. Neither score has much validity for the pilot criterion.

Camouflaged Figures, CP810A [11]

This test is one of the series designed to measure the ability to distinguish a pattern from a confused background.

Description.—The test items consist of capital letters outlined in dots and surrounded by other dots which disrupt the pattern of the letters. The task of the examinee is to distinguish the letter pattern from its confused background. The answers are recorded on a special IBM answer sheet, with spaces for all letters of the alphabet except Q. There is only one letter in each design, and it is always right side up. Three sample items are presented in figure 17.9.

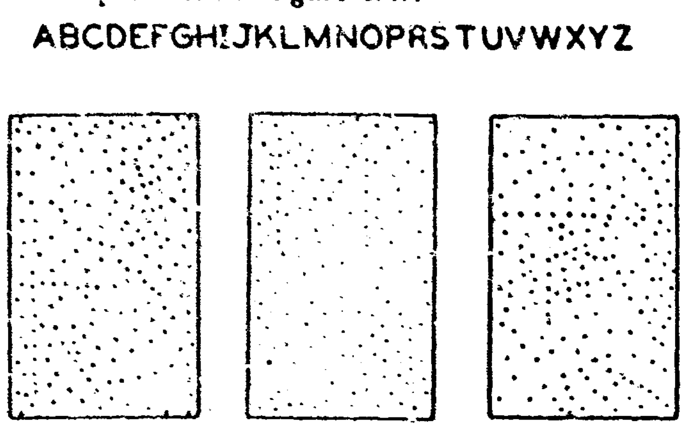


FIGURE 17.9
SAMPLE ITEMS OF CAMOUFLAGED FIGURES, CP810A

[11] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Stuart W. Cook, Lois G. Wright.

(1) *Internal characteristics.*—The test is made up of 2 parts of 30 items each. There are three practice problems at the beginning of the test.

(2) *Administration.*—Thirteen minutes are allowed for part I and 11 minutes for part II. Administration of directions requires approximately 5 minutes.

(3) *Scoring.*—Rights and wrongs are scored separately.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 17.13.

TABLE 17.13.—*Distribution constants for Camouflaged Figures, CP810.!, based upon a sample of pilots in primary training*[1]

Score	N	M	SD
Rights	1,330	32.5	9.2
Wrongs	1,330	12.7	7.6

[1] Tested in March 1944 at Psychological Research Unit No. 3 in class 44I.

(2) *Reliability coefficients.*—Samples to which this test was administered yielded the estimates of reliability given in table 17.14.

TABLE 17.14.—*Estimated reliability coefficients for Camouflaged Figures, CP810A, based upon samples of pilots*

Variables	N	r'_{tt}	r_{tt}
Part I rights *v.* Part II rights	[1]760	0.69	0.82
Part I wrongs *v.* Part II wrongs	[1]775	.40	.57
Part I rights *v.* Part II rights	[2]169	.70	.82
Part I wrongs *v.* Part II wrongs	[2]169	.49	.66

[1] Tested in March 1944 at Psychological Research Unit No. 3. In class 44I.
[2] Tested April 10 and 11, 1944, at Psychological Research Unit No. 3.

(3) *Difficulty.*—Based upon the responses of 733 pilots, the test yielded a mean proportion of correct responses of 0.72, with a range from 0.00 to 0.99 and a standard deviation of 0.20. These statistics are not corrected for chance success, since there is only 1 chance of 25 of securing the correct answer by guessing.

(4) *Test validity.*—Validation results are given in table 17.15.

TABLE 17.15.—*Validity data for Camouflaged Figures, CP810A, based upon samples of pilots in primary training, graduation-elimination criterion*

N	Part	Scoring formula	r_{tt}	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
[2]1,330	I and II	Rights	0.86	32.72	31.36	9.15	0.08	0.17
[2]1,330	I and II	Wrongs	.86	12.70	12.64	7.49	.01	−.05
[3]169	I	Rights	.70	15.64	14.44	4.5	.16	.21
[3]169	I	Wrongs	.70	6.11	6.76	3.5	−.11	−.11
[3]169	II	Rights	.70	15.75	14.56	5.2	.14	.23
[3]169	II	Wrongs	.70	3.92	3.92	2.9	.00	−.02
[3]169	I and II	Rights	.70	31.39	29.00	7.9	.18	.26
[3]169	I and II	Wrongs	.70	10.03	10.68	5.09	−.08	−.09

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44I. Tested in March 1944 at Psychological Research Unit No. 3.
[3] Tested April 10 and 11, 1944, at Psychological Research Unit No. 3.

(5) *Item validity.*—Validation of items revealed a mean phi of 0.01, with a range from −0.11 to +0.24 and a standard deviation of 0.07.

based upon the responses of 600 graduates and 133 eliminees from primary pilot training.

Evaluation.—This test has satisfactory reliability and moderate pilot validity. The average difficulty level of the items is rather low.

Ability to Listen in Noise, CP704BX2

This test was designed to measure the ability to hear instructions through noise. Inability to hear instructions and comments through the intercommunication systems of training airplanes was frequently reported as a difficulty and potential cause of elimination in early phases of training. The first form of this test was developed by the Harvard Psycho-Acoustic Laboratories,[12] and the phonograph records on which this test was recorded were provided by that organization.

Description.—This test, involving the ability to hear spoken words above a noise screen similar to that made by a twin-engine bomber, is recorded on phonograph records, and it involves 20 sections of 10 words each. There are also two practice sections of 10 words each at the beginning of the test. The task of the examinee is to select the word spoken from five alternate choices, and to enter it on his answer sheet. Two sample items are as follows:

	A	B	C	D	E
41.	hurt	church	perch	first	none of these.
42.	platter	clatter	clutter	flatter	none of these.

For 41, the spoken word was "hurt"; for 42, the spoken word was "flatter." The alternative words are the most popular incorrect responses that were made by a group of 100 unclassified aviation cadets who wrote down what they heard.

Part I consists of 20 practice items and 100 scored items; part II consists of 100 scored items.

(1) *Administration.*—The examinees are told that: "This is an auditory test of your ability to hear and identify words above the roar of an airplane engine. * * * As each word is spoken, you will look over the five possible answers * * * and select the word you heard spoken. If the spoken word does not appear, your answer will be E or none of these." The total testing time is approximately 35 minutes.

(2) *Scoring.*—The scoring formula is R—W/5.

Statistical results. (1) *Distribution statstics.*—Table 17.16 presents distribution data for this test.

[12] First form developed by Drs. S. S. Stevens and L. L. Beranek. The records and accessory test materials were kindly made available to the Army Air Corps by Dr. Stevens for experimental use.

TABLE 17.16.—*Distribution data for Ability to Listen in Noise, CP704BX2, based upon samples of classified pilots*

N	Part	Score	M	SD
[1]610	I	R−W/5	45.3	9.5
[2]500	I	R	54.3	8.4
[2]500	II	R	51.5	10.5
[2]500	I and II	R	105.9	16.4
[2]500	I	W	44.2	8.0
[2]500	II	W	47.1	9.6
[2]500	I and II	W	91.3	15.3
[2]704	I and II	R−W/5	94.6	17.8
[3]1,122	I	R−W/5	44.2	9.9

[1] In class 44H, tested at Psychological Research Unit No. 3 in February 1944.
[2] Tested at Psychological Research Unit No. 3 in June 1944.
[3] In classes 44H and 44I, tested at Psychological Research Unit No. 3, February and June 1944.

(2) *Reliability coefficients.*—Samples to which this test was administered yielded the estimates of reliability given in table 17.17. The low correlations between halves of part II is due to the inferior quality of the first side of the second record. In the sample of 500, this defect was remedied.

TABLE 17.17.—*Alternate-forms reliability coefficients for Ability to Listen in Noise, CP704BX2, based upon samples of classified pilots*

N	Variables	r'_{tt}	r_{tt}
[1]403	Part I *v.* Part IIA[2]	0.28	0.44
[1]403	Part I *v.* Part IIB	.43	.60
[1]1,564	Part I *v.* Part II	.51	.68
[1]403	Part IIA[2] *v.* Part IIB	.28	.44
[3]500	Part I *v.* Part II (rights only)	.46	.63
[3]500	Part I *v.* Part II (wrongs only)	.51	.67

[1] Class 44H, tested at Psychological Research Unit No. 3, February 1944.
[2] Side A of part II was technically inferior.
[3] Class 44I, tested at Psychological Research Unit No. 3, June 1944.

(3) *Correlation between rights and wrongs.*—For a sample of 500 pilots (class 44I, tested in June 1944 at Psychological Research Unit No. 3), the correlation between rights and wrongs was −0.88.

(4) *Difficulty.*—Based upon analysis of the responses of 750 pilots, the test yielded a mean proportion of correct responses of 0.57, corrected for chance success, with a range from 0.00 to 0.95 and a standard deviation of 0.22.

(5) *Test validity.*—Validation results are presented in table 17.18.

TABLE 17.18.—*Validity data for Ability to Listen in Noise, CP704BX2, based upon samples of pilots in primary training, graduation-elimination criterion*

Part	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_{c}r_{bis}$[1]
I[2]	[3]610	0.82	45.5	44.2	9.5	0.07	0.10
I and II[4]	[5]704	.84	95.10	92.33	17.80	.09	.12
I[2]	[5]1,122	.87	44.28	43.58	9.86	.04	.07

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Part II omitted because a large proportion of examinees was subjected to technically inferior record.
[3] In class 44H, tested at Psychological Research Unit No. 3, February 1944.
[4] Both parts properly administered.
[5] In classes 44H and 44I, tested at Psychological Research Unit No. 3, February and June 1944.

703320—47—29

(6) *Item validity.*—Validation of items revealed a mean phi of 0.01, based upon the responses of 750 graduates and 68 eliminees from primary pilot training in class 44H. The standard deviation of phi values was 0.08, and the range was from −0.18 to 0.23.

(7) *Study of the effects of seating.*—The effect of the acoustics in the testing room was considered of potential importance as a factor influencing scores on the test. The test was administered in two different buildings. A sample of 308 unclassified aviation cadets showed a mean score of 42.8 and a standard deviation of 9.7 for part I in building A. A sample of 159 unclassified aviation cadets showed a mean score of 46.4 and a standard deviation of 9.2 for part I in building B. Although the difference between the means of the two buildings is significant at the 1 percent level, consideration of the wide range of scores indicated that there would be little effect upon the validity of the test.

It was also found that building A had less standard acoustical conditions than B. For statistical analysis, the room in building A was divided into quadrants. It was found that the left rear quadrant showed a mean of 39.5 and a standard deviation of 8.4, which is significantly lower (at the 1 percent level) than the average of the other three (mean of 43.5, and standard deviation of 9.8). No differences among the quadrants were found for building B. Building A had only three speakers, as opposed to building B with four.

Variations.—There are two preliminary forms of this test, CP704A and CP704BX1. All the forms use the same records, differing only in answer booklets and directions. As was mentioned above, the A form was developed by the Harvard Psycho-Acoustic Laboratories. The directions and booklet for this test require the examinee to write the word that he heard. This test was administered to 100 unclassified aviation cadets, and the most frequently appearing incorrect answers were selected for use in the construction of the first multiple choice form, CP704BX1. The use of this form necessitated extensive changes in the administrative directions. With the exceptions of minor changes in the directions and misleads, form BX2 is identical with BX1.

Evaluation.—Performance on this test is apparently easily influenced by environmental conditions. Both the acoustics of the test room and the quality of the recording must be standardized. Under conditions of large-scale administration, the test has moderate reliability. Because of its negligible pilot validity, however, further development of the test was not undertaken.

ILLUSIONS TESTS

The development of these tests was prompted by two reasons, one general, and one specific.

The specific reason was that several valid tests of size and distance estimation seemed to be contaminated with the presence of illusory effects, e. g., the Estimation of Length test (see ch. 18). It was felt desirable, therefore, to construct separate tests that would attempt to measure resistance (or, conversely, susceptibility) to illusions, in a deliberate and systematic manner.

In general, it was thought desirable to undertake an extensive investigation of objectivity of perception, i. e., correspondence of the perceived dimensions of objects with physically measured dimensions. The project was begun quite late in the program, however, and only few data are available.

Objectivity of Perception, CP806CX1 and CP806CX2 [11]

These two tests both utilize familiar geometric illusions, such as the Müller-Lyer, Poggendorf, Ponzo, Sanders parallelogram, vertical-horizontal, filled *v.* unfilled space, equal squares, Titchener's circles, and unnamed variants of these.

(1) *Internal characteristics.*—Form CP806CX1 consists of 80 items. Five types of illusions are represented by 10 items each, and 2 other illusions by 15 items each. Form CP806CX2 contains 70 items, with each of 7 illusions represented by 10 items. The 14 illusions in the 2 forms are all different, but 1 form may contain a major variant of the other. Each illusion is presented as a separate section, including all the items for that illusion. In figure 17.10 are shown items representative of several of the illusions. For illustrative purposes, the items shown are those in which the indicated dimensions are equal.

Each set of items constitutes a series, established as satisfactory by pretesting. For example, for the pan illusion illustrated in the panel I of figure 17.10, there are two standard lengths, 1½ and 2 inches. For variables there are lines of the following lengths: 1½, 1⅝, 1 6/8, 1⅞, and 2 inches for the first standard; and 2, 2 3/16, 2 4/16, 2 6/16, and 2 8/16 for the second standard. The order of presentation of the items is randomized with respect to length of variables and position (right or left) on the page of the variables.

(2) *Administration.*—The directions to the test are designed to inculcate a set to resist the illusions, and, at the same time, to work very rapidly. Pertinent parts of the directions follow:

> This is a test of your ability to detect rapidly, merely by inspection, the true sizes of camouflaged figures. * * *
>
> Some of these figures have been drawn to look different than they really are. That is, a line in one of these drawings may look longer or shorter than a ruler would show it to be * * *
>
> Your task is to make the most accurate judgments you can, by attempting to ignore the extra lines and angles which camouflage the true size of the various parts of the figures * * *
>
> Work rapidly. The use of artificial aids is strictly to be avoided. As a matter of fact, if you stop to measure length of lines or to turn the figures around, for example, you will not be able to finish the test * * *

[11] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. John I. Lacey, Lt. F. A. Lipman, and Sgt. Albert H. Hastorf III.

To reinforce the set for speed, the test administrator paces the examinees, announcing at the end of each minute on what item the examinee should be working. The examinees are required to answer eight items each minute. This pacing was established after experimental administration of the tests to 500 unclassified aviation students at Psychological Research Unit No. 3. It permits approximately 80 percent to finish all items and 100 percent to finish all but two items in a section.

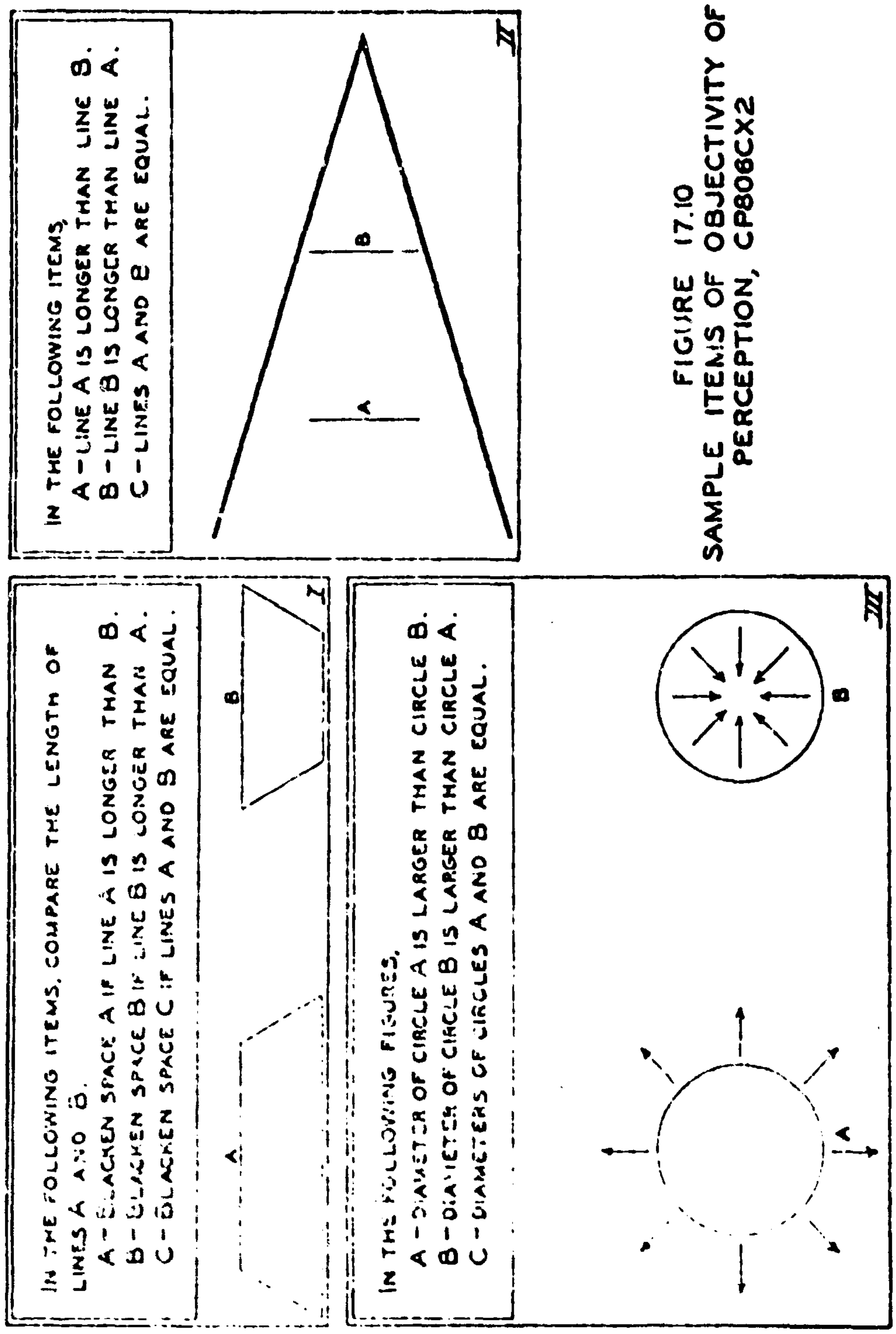


FIGURE 17.10
SAMPLE ITEMS OF OBJECTIVITY OF PERCEPTION, CP808CX2

(3) *Scoring.*—No scoring formula is yet recommended for these tests. For experimental purposes, responses were categorized as indicating maximum resistance to illusion (i. e., calling equal lines equal despite illusory effects), maximum susceptibility to illusion (i. e., answering in the direction of the illusion), and intermediate judgments (e. g., two lines being called equal, where maximum resistance to the illusion would elicit a response of longer, and maximum susceptibility a response of shorter).

Statistical results. (1) *Distribution statistics.*—Available distribution data are shown in table 17.19.

TABLE 17.19.—*Distribution data and Kuder-Richardson reliability for Objectivity of Perception, CP806CX1 and CP806CX2, based upon samples of unclassified aviation students*

N	Form	Score	M	SD	r_{tt}[1]
[2]551	CP806CX1	"R"[3]	23.5	6.6	0.66
[4]500	CP806CX1	"R"	22.1	6.8	...
[2]551	CP806CX1	"W"[5]	21.1	5.9	.59
[4]500	CP806CX1	"W"	19.8	5.8	...
[2]551	CP806CX1	"I"[6]	18.8	5.9	.50
[2]551	CP806CX2	"R"	34.9	6.1	.58
[4]500	CP806CX2	"R"	33.8	6.4	...
[2]551	CP806CX2	"W"	14.2	4.5	.46
[4]500	CP806CX2	"W"	14.1	5.2	...
[2]551	CP806CX2	"I"	15.8	5.8	.66

[1] Using Kuder-Richardson formula No. 21.
[2] Tested February through April 1945, at Psychological Research Unit No. 2.
[3] "R" means responses indicating resistance to illusions.
[4] Tested April and May 1945 at Medical and Psychological Examining Unit No. 8.
[5] "W" means responses indicating susceptibility to illusions.
[6] "I" means intermediate judgments, as defined in the text.

(2) *Reliability coefficients.*—Although the assumptions underlying the formula are not completely satisfied, Kuder-Richardson estimates (formula No. 21) were secured as preliminary evidence concerning test reliability. The data are presented in the last column of table 17.19. Correlations between the two forms also yield some indication of the reliabilities of the test, although the two forms are perhaps not comparable. The data are shown in table 17.20.

TABLE 17.20.—*Correlations between Forms CP806CX1 and CP806CX2 of Objectivity of Perception, based upon a sample of 551 unclassified aviation students*[1]

Variable	1	2	3	4	5	6
1. "R" for CP806CX1	...	−0.37	−0.30	0.42	−0.18	−0.30
2. "W" for CP806CX1	−0.37	...	−.06	−.27	.44	−.02
3. "I" for CP806CX1	−.30	−.06	...	−.31	−.21	.52
4. "R" for CP806CX2	.42	−.28	−.31	...	−.40	−.72
5. "W" for CP806CX2	−.18	.44	−.21	−.40	...	−.27
6. "I" for CP806CX2	−.30	−.02	.52	−.72	−.27	...

[1] Tested February through April 1945 at Psychological Research Unit No. 2.

Evaluation.—The test appears to be quite difficult and the three scores on each test appear to have only moderate reliability. The moderate correlations (0.42 to 0.52) between corresponding scores on the two forms indicate either relative lack of common-factor variance, and thus a tendency towards specificity of susceptibility to different illusions, or low reliability.

Normality of Perception, CP806CX3 and CP806CX4 [14]

These two tests utilize the same test booklets as Objectivity of Perception, CP806CX1 and CP806CX2, respectively. The sole difference lies in the instructions to the test. In taking an illusions test, the examinee may adopt either a naive, phenomenological approach, or an approach characterized by the intent to see objects as they are physically. It was decided to investigate these two approaches, to discover what differences in test performance might be attributable thereto, and which mental set, if either, produced the more reliable and valid results.

(1) *Administration.*—The instructions to these tests attempt to inculcate a naive, nonresistant set. Pertinent extracts from the directions follow:

> In this test you will be asked to report things exactly as you see them * * * Many figures have been drawn deliberately to look different than they actually are. That is, a line in one of these drawings may look longer or shorter than a ruler would prove it to be * * * Your task is to report how the figure looks to you, not how you think it should look * * * You will get the best score by working rapidly and recording your first impression of how the figures actually look to you.

Statistical results. (1) *Distribution statistics.*—Distribution data are shown in table 17.21.

TABLE 17.21.—*Distribution data and Kuder-Richardson reliabilities of Normality of Perception, CP806CX3 and CP806CX4, based upon a sample of 518 unclassified aviation students*[a]

Form	Score[b]	M	SD	r_{tt}[c]
CP806CX3	"R"	21.2	5.9	0.60
CP806CX3	"W"	25.3	6.8	.67
CP806CX3	"I"	16.6	5.2	.55
CP806CX4	"R"	31.3	5.2	.40
CP806CX4	"W"	18.2	5.1	.50
CP806CX4	"I"	14.6	5.4	.62

[a] Tested in February and March 1945 at Psychological Research Unit No. 2.
[b] For meaning of scores, see footnotes to table 17.19.
[c] Using Kuder-Richardson formula No. 21.

(2) *Reliability coefficients.*—Kuder-Richardson reliabilities were computed for these forms also, and are shown in the last column of table 17.21. Again, correlations between the forms yield some indication of test reliability and they are presented in table 17.22.

[14] Developed at Psychological Research Unit No. 3. Chief contributor: Capt. John I. Lacey.

TABLE 17.22.—*Correlations between Forms CP806CX3 and CP806CX4 of Normality of Perception based upon a sample of 518 unclassified aviation students*[1]

Variable	1	2	3	4	5	6
1. "R" for CP806CX3	...	−0.32	−0.12	0.32	−0.19	−0.13
2. "W" for CP806CX3	−0.32	...	−.27	−.24	.51	−.19
3. "I" for CP806CX3	−.12	−.27	...	−.15	−.31	.46
4. "R" for CP806CX4	.32	−.24	−.15	...	−.33	−.49
5. "W" for CP806CX4	−.19	.51	−.31	−.33	...	.51
6. "I" for CP806CX4	−.13	−.19	.46	−.49	−.51	...

[1] Tested in February and March 1945 at Psychological Research Unit No. 2.

Evaluation.—The illusions tests administered with instructions intended to produce a naive, phenomenological set do not seem to differ much from the previous forms. Reliabilities and score intercorrelations seem fairly comparable. Administering the tests with normality instructions rather than objectivity instructions results in fewer judgments indicating resistance to illusions and more judgments indicating susceptibility to illusions (compare tables 17.19 and 17.21), but the differences are quite small.

Objectivity of Perception, CP806BX1, and Normality of Perception, CP806BX3 [15]

These tests are comparable in arrangement, format, and directions to the previous forms. They differ, however, in content, being devoted to illusory effects stressed by Gestalt psychologists, centering about the concepts of structure and articulation of parts. The two sets of instructions are entirely comparable to those used in the previous tests.

The test booklet is divided into 10 parts, comprising 125 items. Sample items of three different sections are shown in figure 17.11. For illustrative purposes, the correct answers to the items chosen for reproduction are all "equal." The nature of the test may be made clear in the following listing which gives the phenomenon of perception with which several of the parts are concerned.

Part I: Parts of well-structured objects appear less numerous than irregularly spaced parts of non-structured groups (see upper panel of fig. 17.11).

Part II: Parts of homogeneous regular objects appear less numerous than parts of broken, irregular objects (see middle panel of fig. 17.11).

Part III: Articulated groups composed of identical objects (plane silhouettes) appear less numerous than unorganized groups composed of dissimilar objects (see lower panel of fig. 17.11).

Part IV: Triangles appear larger in area than Greek crosses.

Part VII: Of two equilateral triangles, one placed above the other, the lower one is perceived as of lesser area.

Part IX: A tilted square is perceived as of greater area than the same square presented in the horizontal-vertical orientation.

Other parts constitute variants and combinations of these.

[15] Developed at Psychological Unit No. 3. Chief contributors: Sgt. Stanley Blumberg, Capt. John I. Lacey, and Lt. Eli A. Lipman.

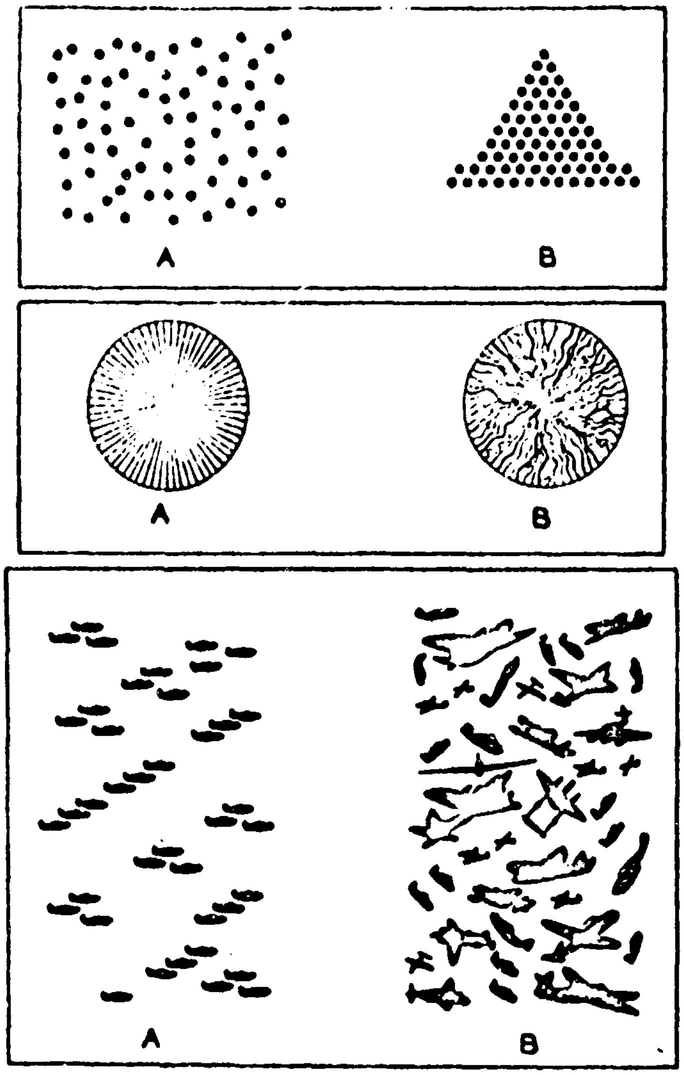


FIGURE 17.11
SAMPLE ITEMS OF OBJECTIVITY OF PERCEPTION, CP806BX1, AND NORMALITY OF PERCEPTION, CP806BX3

Each section of the test is separately timed with 1 minute and 10 seconds allotted to part V, 1 minute and 25 seconds to part IV, 1 minute and 30 seconds to part III, 1 minute and 35 seconds to parts II and IX, and 1 minute and 40 seconds to parts I, VI, VII, VIII and X.

These time limits are based upon preliminary experimental administration of the test to 248 unclassified aviation students at Psychological Research Unit No. 2.

No other statistical data were available at the time this was written.

EVALUATION OF FORM PERCEPTION TESTS

Too few data are available to permit a thorough evaluation of form-perception tests. The lack of extensive factorial data, in particular, obviates the type of interpretation and evaluation that has proved to be so useful and fruitful.

The two pattern-formation tests for which data are available (Picture Integration, CP104A, and Pattern Assembly, CP804A) have moderate pilot validity, attributable in large part to saturations in the space-relations, perceptual-speed, visualization, and length-estimation factors. It is noteworthy that for both these tests there are sizeable discrepancies between communalities and reliabilities. It may be that at least some of this disparity is due to a new factor of perceptual integration. While there are no data to support this speculation, the problem is worthy of study.

The two pattern-completion tests (Mutilated Words, CP512A, and Object Completion, CP811A) apparently have little or no pilot validity, which argues for a lack of the factors known to be valid for pilots. The additional fact that very similar tests were found by Thurstone (2) to have prominent loadings in a factor tentatively identified as "speed and strength of closure" implies that this factor lacks pilot validity.

The pattern-analysis tests promise only low to moderate pilot validity. The value of the illusions tests is yet unexplored. Available data indicate relative specificity of the various illusions, although Thurstone (2) has tentatively established the existence of a factor common to different geometric illusions.

BIBLIOGRAPHY

(1) Street, R. F. *A Gestalt Completion Test*, Columbia University, 1931.

(2) Thurstone, L. L. *A Factorial Study of Perception*, University of Chicago Press, 1944.

CHAPTER EIGHTEEN

Size and Distance Estimation Tests[1]

INTRODUCTION

Contents of the Chapter

This chapter considers tests that measure the ability to perceive size and distance accurately. For convenience, not because of a considered psychological analysis, the tests are divided into three categories, to each of which a section is devoted. The categories are: (1) Distance Judgment Tests, (2) Angular Judgment Tests, and (3) Judgment of Proportions Test. Following the discussion of the tests, a general evaluation of the area is presented.

Rationale for Development of Size and Distance Estimation Tests

Piloting a plane, navigating a plane, sighting on a target—all these activities seem to involve accurate perception of size and distance. Early job-analysis data provided some indication of the importance of this perceptual activity in each of the three main air-crew positions.

In table 1.5, it may be seen that estimation of speed and distance was mentioned in from 27 percent to 38 percent of eliminations from elementary and advanced training, and in 1 percent (in one sample) and 5 percent (in another sample) of operational reclassifications.

When combat supervisors rated 20 traits for their importance in combat, on a 9-point scale in which the nominal figure 5 was taken to mean "better than average," estimation of speed and distance was rated 5.8 for combat bombardiers (see table 1.2), 6.6 for combat navigators (see table 1.4), 7.5 for combat fighter pilots, and 6.1 for combat bomber pilots (see table 1.6).

Late in the program (February 1944), an intensive job analysis of the act of landing a plane was undertaken by Psychological Research Unit No. 3, which supplied some very specific information on the role of distance perception. Four aviation psychologists spent 3½ weeks at a primary school collecting data on individually observed landings, interviewing students and instructors, and themselves experiencing landings. The main impetus for this research was provided by the fact that a large number of eliminations from elementary training occurred during the period of concentrated practice on landings.

[1] Written by Capt. John I. Lacey and Tech/Sgt. Gerald H. Shurley.

While the job-analysis data cannot be presented in detail here, some description of the role of distance estimation may be given. In turning into his approach to the landing lane, the pilot must accurately judge distance so that he may cut the throttle and make the gliding turn at the correct time. As the airplane nears the ground, the pilot causes the plane to lose air speed until it "stalls out" and drops to the ground. A perfect landing can occur only when the pilot accurately judges his height off the ground. On the ground, of course, the pilot must estimate the distance of other objects from his own plane.

Data such as these more than justify the development of tests of distance estimation, when the distances involved are relatively great. Implicit in the development of printed tests of length estimation, however, is the assumption of a factor of length estimation common to a wide range of distances. A similar assumption is made in the development of printed tests of judgment of angular magnitudes.

DISTANCE JUDGMENT TESTS

In this section will be described those tests that measure the ability to estimate linear and nonlinear extents.

Shorter Line, CP606

Ea n the classification program, the best available information concerning the duties of pilots and navigators indicated that tests sampling the abilities to estimate size and distance, to make quick and accurate approximate readings from tables, graphs, and meters, to apprehend quickly number-size, etc., had promise of validity. These diverse activities were grouped together under the term "Quantitative Perception," and the Office of the Air Surgeon, Psychological Branch, requested the Cooperative Test Service to supply a series of such tests. In this chapter, three of these tests, Shorter Line, CP606, Nearest Point, CP607, and Shortest Path, CP608, are discussed.[3]

Shorter Line, CP606, is part VI of the Quantitative Perception Test. This test was used in the classification battery of April 1942.

Description.—In this test, the examinee is required to indicate which of two lines is the shorter.

(1) *Internal characteristics.*—The test comprises 30 items which are printed on one-half of one side of an IBM answer sheet. Each item consists of five straight lines of different lengths radiating from a central point. Distracting lines, curves, and figures are added. Two of the lines are labeled with the letter-symbols A and B. The task of the examinee is to select the shorter line of the two that are labeled, and to record his answer in the space to the right of the diagram. Two examples of items are shown in figure 18.1.

[3] For a discussion of the other parts of the Quantitative Perception Test, see ch. 16.

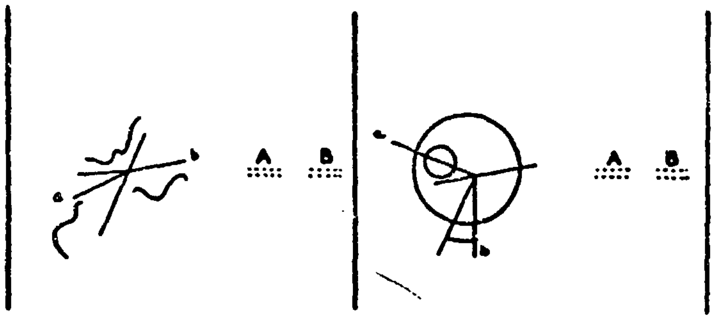


FIGURE 18.1
SAMPLE ITEMS OF SHORTER LINE, CP606

(2) *Administration.*—Administrative directions, greatly simplified by having the test items on the answer sheet, are quite short. They state:

In each of the figures below, there are five straight lines radiating from a central point. One line is labeled *a* and another *b*. Decide which of these two lines is SHORTER, and blacken the appropriate answer-space below at the right of the figure. Pay no attention to any lines through or near the five "spokes" radiating from the central point . . . You will have 2 minutes for this part . . .

(3) *Scoring.*—The scoring formula used in the classification battery was R−3W.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 18.1. The distribution curves are approximately symmetrical.

TABLE 18.1.—*Distribution constants for Shorter Line, CP606*

Group	N	M	SD
Unclassified aviation students[1]	240	22.1	5.9
Unclassified aviation students[1]	527	21.6	6.2
Navigators[2]	392	21.9	6.1

[1] Testing dates and units unidentified.
[2] In Hondo classes 42-11 to 42-16. Tested at Psychological Research Unit No. 2.

(2) *Factorial composition.*—The most significant loading (0.44) is in the length-estimation factor in which the test is almost pure. This loading is the weighted average of the loadings in two analyses. The communality was found to be very low (0.27). For a fuller picture of the factorial composition of this test, see appendix B.

(3) *Test validity.*—Validation results based on several samples are given in table 18.2.

Evaluation.—Shorter Line shows only moderate pilot and navigator validity. There are no data concerning the reliability of the test, but the evidence of the low coefficients for Nearest Point, CP607, and Shortest Path, CP608, suggest that it too is quite low.

Factor analyses of this test show that 27 percent of the total variance has been accounted for by common factors. Of this, only the length-estimation factor accounts for a significant amount (19 percent) of the total variance of the test. The remaining 8 percent is accounted for by

TABLE 18.2.—*Validity data for Shorter Line, CP606*

Group	Class	Criterion	N_t	p_g	M_g	M_e	SD_t	r_{bis}
Pilots in primary training	42G	Graduation-elimination	729	0.70	20.71	18.97	...	0.17
Pilots in primary training	42F to 42H	do	676	.59	20.57	18.60	...	.19
Navigators[1]	42-11 to 42-16	do	163	.71	21.65	22.10	5.71	−.05
Do[2]	42-11 to 42-16	do	392	.77	22.41	20.57	5.70	.19
Bombardiers	42-11 to 42-16	Combat circular error[3]	307	...	....	....	...	[4].05

[1] Tested at Psychological Research Unit No. 1. New aviation students.
[2] Reclassified pilots, tested at Psychological Research Unit No. 2.
[3] A measure of bombing accuracy obtained during training. A highly unreliable criterion.
[4] Product-moment coefficient.

factors on which the loadings are quite low. While Pattern Assembly, CP804A (see ch. 17), has a higher loading on the length-estimation factor, Shorter Line is a purer measure of this factor. The loading is sufficiently high to suggest that, with an increased reliability, the test may find value in future factor research as well as in selection where this factor is valid. It is likely that the distracting lines merely serve to create illusions and to render the test more ambiguous factorially. In a later test of the same kind, distractors were omitted.

An estimate of the pilot validity of this test, made from factorial information (see ch. 28) is 0.11, which is somewhat short of the empirical validity. This discrepancy may be large enough to indicate some validity in other factors that were not included in this estimate. Other possible explanations of the discrepancy would be underestimation of the length-estimation factor loading in the test CP606, or in the pilot criterion, or in both.

Nearest Point, CP607

This is part VII of the Quantitative Perception Test. This test was also in the classification battery of April 1942.

Description.—Like Shorter Line, this test is printed on one-half of one side of an IBM answer sheet and is so designed that there are but two alternative answers.

(1) *Internal characteristics.*—Each item consists of five dots irregularly scattered around a reference point, which is in the form of a dot with a small circle around it. Two of the dots are labeled *a* and *b*. The task of the examinee is to select, from the labeled dots, the one that is nearer the reference point. In many of the items, lines, curves, and figures are drawn in and around the pattern formed by the dots. There are 30 items in the test. Two illustrative items are shown in figure 18.2.

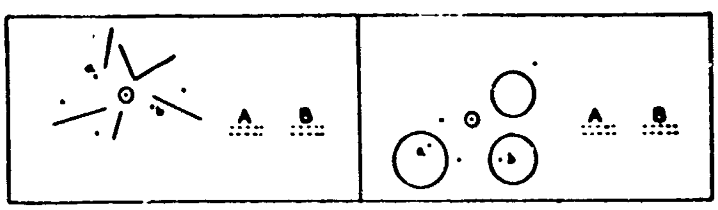


FIGURE 18.2
SAMPLE ITEMS OF NEAREST POINT, CP607

(2) *Administration.*—One completed sample item is given in the directions in explaining the test. Two minutes are allowed for testing time. The directions instruct the examinee to work carefully, and that each error will result in a deduction of three points from the score.

(3) *Scoring.*—The scoring formula is R−3W.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 18.3. The distribution curves are approximately symmetrical.

TABLE 18.3.—*Distribution constants for Nearest Point, CP607*

Group	N	M	SD
Unclassified aviation students[1]	527	15.9	6.9
Unclassified aviation students[1]	243	17.0	6.8
Navigators[2]	392	17.6	7.3

[1] Testing dates and units unidentified.
[2] In Hondo classes 42-11 to 42-16. Tested at Psychological Research Unit No. 2.

(2) *Reliability coefficient.*—By the odd-even method an estimated reliability coefficient of 0.87, corrected for length, was obtained. This figure is based on a sample of 243 unclassified aviation students. Since the test is speeded, this figure is an overestimation. Because the scoring formula (R−3W), however, emphasizes errors, the overestimation may not be serious.

(3) *Factorial composition.*—The most significant weighted-average loadings are in the length-estimation (0.43), general-reasoning (0.21), and visual-memory (0.23) factors. The loading in general reasoning is suspect. The communality is 0.38. For a fuller picture of the factorial composition of this test, see appendix B.

(4) *Test validity.*—Validation results based on several samples are given in table 18.4.

Evaluation.—This extremely simple test has good pilot validity (average coefficient approximately 0.19 based upon 4,045 cases). Since the test is only 2 minutes long, however, its reliability is low. Lengthening would probably improve both reliability and validity at low cost in time. The estimated validity, utilizing factorial information (see ch. 28), is 0.14, which is a little short of the empirical validity.

In the factor analysis of this test, 38 percent of the total variance was accounted for by common factors. The length-estimation factor accounts for 18 percent, the general-reasoning factor for 4 percent, and the visual-memory factor for 5 percent of the total variance. The remaining 11 percent is accounted for by factors in which the loadings are quite low.

Shortest Path, CP608

This is part VIII of the Quantitative Perception Test, also included in the classification battery of April 1942.

Description.—The test consists of 30 items printed on one-half of one side of an IBM answer sheet.

(1) *Internal characteristics.*—Each item consists of two points, labeled P and Q, placed about 1½ inches apart. Three curved or angular lines, labeled A, B, and C, are drawn between these two points. The task of the examinee is to select the path, or line, between the two points that

TABLE 18.4.—*Validity data for Nearest Point, CP607*

Group	Class	Criterion	N_t	p_s	M_p	M_e	SD_t	r_{bis}
Pilots in primary training	42I to 42K; 43A	Graduation-elimination	726	0.70	15.88	12.63	....	0.28
Do	42G	....do	676	.59	17.09	13.46	....	.31
Do	42H and 42I	....do	1,323	.65	16.71	15.31	7.23	.12
Bombardiers	42-11 to 42-16	Circular error	180	...	....	....	....	[1].03
Do[2]	42-11 to 42-16	....do	307	...	....	....	....	[1].01
Navigators[3]	42-11 to 42-16	Graduation-elimination	163	.71	19.04	17.08	6.49	.18
Do[3]	42-11 to 42-16	....do	392	.77	17.77	16.62	6.79	.10

[1] Product-moment correlation.
[2] Reclassified pilots.
[3] New aviation students.

703320—47—30

is shortest. The examinee records his answer directly below the figure. As illustrations, figure 18.3 presents two items of the test.

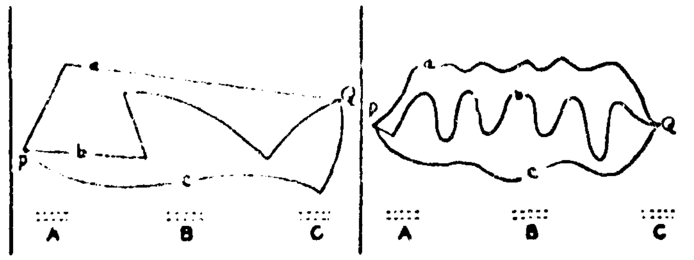


FIGURE 18.3
SAMPLE ITEMS OF SHORTEST PATH, CP608

(2) *Administration.*—One completed problem is shown next to the directions. The testing time is 2 minutes. The examinee is warned to work carefully, since a heavy penalty (three points) is given to errors.

(3) *Scoring.*—The scoring formula is R−3W.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 18.5. The distribution curves are approximately symmetrical.

TABLE 18.5.—*Distribution constants for Shortest Path, CP608*

Group	N	M	SD
Unclassified aviation students[1]	527	17.4	7.3
Unclassified aviation students[1]	223	19.5	7.5
Navigators[2]	392	18.9	7.6

[1] Testing dates and units unidentified.
[2] In Hondo classes 42-11 to 42-16. Tested at Psychological Research Unit No. 2.

(2) *Reliability coefficient.*—By the odd-even method, an estimated reliability coefficient of 0.69, corrected for length, was obtained, based on a sample of 233 unclassified aviation students. This figure is an overestimation, since the test is highly speeded. The scoring formula weights errors heavily, however, so the overestimation may not be serious.

(3) *Factorial composition.*—The most significant loadings are in the length-estimation (0.46), spatial-relations (0.32), visualization (0.28), and perceptual-speed (0.25) factors. The communality is 0.52. For a fuller picture of the factorial composition of this test, see appendix B.

(4) *Test validity.*—Validation results based on several samples are given in table 18.6.

Evaluation.—This test has good validity for pilots (average validity coefficient approximately 0.25) and navigators (average validity coefficient approximately 0.28). Its validity for bombardier is difficult to evaluate because of the low reliability of circular error as a measure of proficiency for that air-crew position.

TABLE 18.6.—*Validity data for Shortest Path, CP608*

Group	Class	Criterion	N_t	p_s	M_s	M_e	SD_t	r_{bis}
Pilots in primary training ..	42D	Graduation-elimination	722	0.70	18.09	15.10	....	.23
Do	42H; 42I	....do	1,323	.65	18.37	15.02	....	.23
Navigators[1]	42-11 to 42-16	....do	163	.71	21.18	18.72	6.77	.22
Do[2]	42-11 to 42-16	....do	392	.77	20.51	16.89	7.16	.29
Bombardiers[1]	42-11 to 42-16	Combat circular error	180	...	....	....	....	[3].02
Do[2]	42-11 to 42-16	....do	307	...	....	....	....	[3].05

[1] New students.
[2] Reclassified pilots.
[3] Product-moment correlation.

Factor-analysis data for this test indicate that 52 percent of the total variance has been accounted for by common factors. Of this, the perceptual-speed factor accounts for 5 percent, the visualization factor 8 percent, the spatial-relations factor 10 percent, and the length-estimation factor 21 percent of the total variance. The remaining 8 percent is accounted for by factors in which the test has very low loadings. It is not as pure a measure as Shorter Line, CP606.

The estimated pilot validity coefficient (computed from factor equations, see ch. 28) is the same as that found empirically. This indicates that all the factors valid for pilot selection are accounted for in this test.

Map Distance, CP626B [9]

The known validity of Nearest Point, CP607, provided the basis for the construction of map distance. It was desired to construct a longer test to increase reliability and to add more face validity to a test of distance estimation. The test was designed also to utilize somewhat longer distances than did the nearest point test.

Description.—The test utilizes four copies of a given portion of an airways map (6½ inches x 7½ inches). A copy of the map appears on each of four pages, arranged in two double-spreads. Twelve towns on each map are indicated by a large dot and a letter-symbol. The 12 towns indicated vary from map to map. A reference point, identified by a dot with a circle around it, is also placed on the map. The position of this reference point also varies from map to map. The task of the examinee is to indicate which of two towns is closer to the reference point. In figure 18.4 is shown one page of the test.

(1) *Internal characteristics.*—As described above, there are 48 items, divided into 2 parts, in this test. A sample map with three items is used at the beginning of the test.

(2) *Administration.*—Pertinent parts of the directions for the test are as follows:

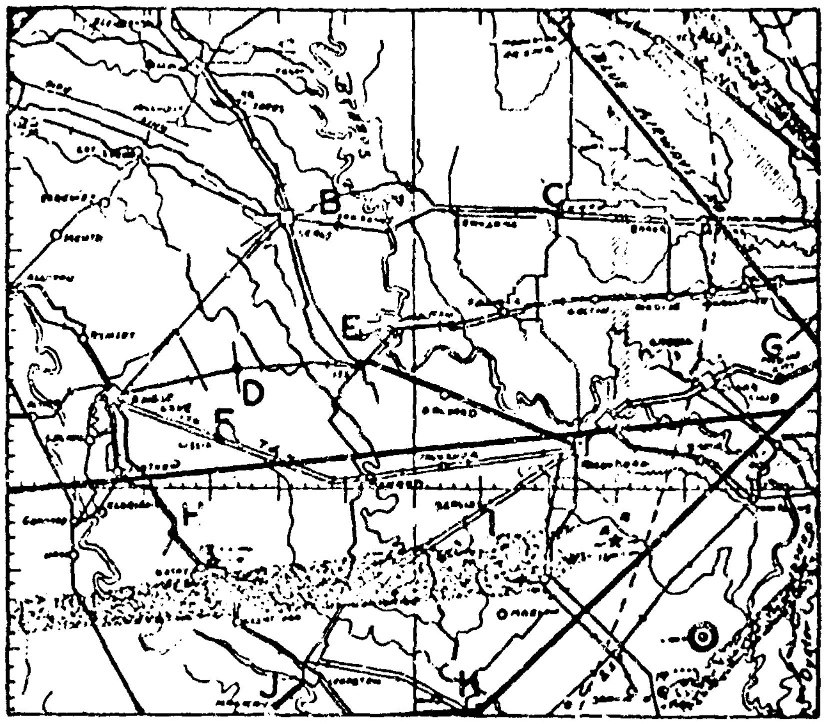
> This is a test of your ability to estimate distances on maps. Suppose that you are a pilot and your problem is to judge which of two towns is the closer. In order to determine this, you will have to consult your map and estimate the distance from your plane to each of the towns in question. In this test, you will be shown a map and asked to compare the distances from a given reference point to various pairs of towns. Your task each time will be to choose the nearer town.

Each double-spread is separately timed. Three minutes are allowed for each part and 5 minutes for the administration. For experimental purposes, however, 5 or 9 minutes were occasionally allowed for each part.

(3) *Scoring.*—The scoring formula is $R-3W+40$.

[9] Developed at Headquarters, AAF Training Command. Chief contributors: Maj. S. Rains Wallace and Staff of Perceptual Research Unit.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 18.7. The distribution curves are approximately symmetrical and considerably flatter than normal.



INDICATE WHICH ONE OF THE POINTS IN EACH PAIR LISTED BELOW IS NEARER TO THE REFERENCE POINT IN THE LOWER RIGHT-HAND CORNER OF THE MAP.

1-G OR I	5-K OR G	9-H OR D
2-D OR F	6-J OR C	10-H OR F
3-E OR J	7-F OR B	11-C OR E
4-B OR D	8-I OR K	12-B OR H

FIGURE 18.4
SAMPLE MAP & ITEMS OF MAP DISTANCE, CP626B

TABLE 18.7.—*Distribution constants for Map Distance, CP626B, based on a sample of 315 classified pilots*[1]

Scoring formula	M	SD	Range of scores	
			Low	High
Rights only	16.5	4.7	2	35
R−3W+40	22.5	13.8	−20	53

[1] In classes 44F and 44I. Tested at Psychological Research Unit No. 3.

(2) *Internal consistency.*—Analysis of responses of several sample groups yielded the internal-consistency data given in table 18.8.

TABLE 18.8.—*Internal consistency data for 48 items of Map Distance, CP626B, based upon samples of unclassified aviation students*

Scoring formula	N	Mφ	SDφ	Range of φ's	
				Low	High
R − 3W + 40	[1]750	0.29	0.21	−0.07	0.70
Rights only	[2]417	.19	.10	.03	.37
R − 3W	[2]417	.32	.19	−.06	.74

[1] Tested at Psychological Research Unit No. 3; dates unidentified.
[2] Same sample, tested at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—One sample yielded the estimates of reliability given in table 18.9.

TABLE 18.9.—*Estimated reliability coefficients (odd-even) for Map Distance, CP626B,[1] based upon samples of unclassified aviation Students*

N	Score	r_{tt}[2]
[3]185	R − 3W + 200	0.72
[3]185	R − W + 200	.6[illegible]

[1] With a testing time of 10 minutes.
[2] Note that this is a speeded test, but with heavy penalty for errors. Odd-even reliabilities, therefore, are spuriously high.
[3] Testing dates and units unidentified.

(4) *Correlation between rights and wrongs.*—For a sample of 911 navigators (see table 18.10), the correlation between rights and wrongs was only −0.03. For a sample of 239 pilots tested at Psychological Research Unit No. 3 in September, October and November 1944, the correlation was −0.07.

(5) *Difficulty.*—Based upon item analysis of the responses of 750 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.52, corrected for chance, with a range from 0.00 to 0.89, and a standard deviation of 0.27.

(6) *Factorial composition.*—The most significant weighted average loadings are in the visualization (0.33) and length-estimation (0.30) factors. The communality is 0.35, which is far below the test's probable reliability. For a full picture of the factorial composition of this test, see appendix B.

(7) *Test validity.*—Validation results based on several samples are given in table 18.10.

(8) *Item validity.*—Validation of items for a sample of 376 pilots in classes 44F and 44H, tested at Psychological Research Unit No. 3, yielded a mean phi of 0.05, with a range from −0.13 to +0.20 and a standard deviation of 0.09. In this sample, 88 percent was graduates.

A study in scoring.—From the formula for the correlation of sums it can be easily seen that marked changes in the reliability of a test may occur with the assignment of varying relative weights to right and wrong answers. Similar changes can be brought about in validity coefficients.

TABLE 18.10.—*Validity data for Map Distances, CP626B, based upon graduation-elimination from training*

Group	Class	Scoring formula	N_t	p_e	M_e	M_e	SD_t	r_{bis}	$c r_{bis}$
Pilots in primary training	43I	R − 3W + 200[1]	[2]245	0.75	196.58	191.12	18.95	0.17	....
Do	43I	R − 3W + 200[3]	[4]286	.74	197.33	189.80	14.76	.30	....
Do	44F thru 44I	Rights	[4]402	.89	17.61	16.32	4.58	.15	[5]0.22
Do	44F thru 44I	Wrongs	[4]402	.89	10.79	11.46	4.34	−.09	[5]−.15
Navigators	....	Rights	[6]911	.90	18.68	17.30	4.67	.15	[7].33
Do	....	Wrongs	[6]911	.90	10.14	11.10	4.05	−.12	[7]−.13

[1] With a testing time of 10 minutes.
[2] Tested at Psychological Research Unit No. 1.
[3] With a testing time of 18 minutes.
[4] Tested at Psychological Research Unit No. 3.
[5] Assuming an unrestricted stanine standard deviation of 1.83.
[6] Tested at Selman Field in March 1944, and at Ellington Field and Psychological Research Unit No. 3 in February, 1944.
[7] Assuming an unrestricted stanine standard deviation of 2.00.

These modifications in reliability and validity coefficients are most marked when right and wrong answers measure different functions, i. e., when the correlation between rights and wrongs is low.

Data available for the Map Distance test provide a dramatic example of changing factorial composition of a test score with changing emphasis on right and wrong answers.

The problem was set by the results of an internal-consistency item analysis against the criterion of total score, computed by the formula R−3W+40. Utilizing the highest 27 percent and the lowest 27 percent of a group of 447 unclassified aviation students, the mean phi based on total group was 0.10, whereas the mean phi based on total answered was 0.32. These results were completely the reverse of those to be anticipated for a speeded test. In a true speed test where everyone attempting an item responds correctly and a good score depends merely on the speed of response, the phis based on total answered would all be zero, whereas the phis based on total group would regularly increase from some one item in the test to the last item. As the speed element in a test decreases in importance, the discrepancy between phis computed on the two bases continuously decreases, until, with a pure power test, it disappears.

The results of the item analysis of the Map Distance test indicate, then, that this test, scored to weight errors heavily, would be more internally consistent if it were administered as a power test. Further investigation of the problem was indicated. Utilizing the same sample of 447 individuals, a new item analysis was performed against the criterion of total rights only. The results were in accord with expectation; the mean phi based on the total group was now 0.31, and that based on total answered was 0.18. Apparently, then, the rights score is a speed score, whereas the error score is a power score.

These results led to a small but revealing intercorrelational study. Using a sample of 315 unclassified aviation students and scoring the papers with two formulas, R and R−3W+40, correlations with a selected group of tests were computed. The tests were selected to reveal possible changes in factorial composition. The data are presented in table 18.11.

TABLE 18.11.—*Correlations of two scores on Map Distance, CP626B, with selected tests*

Test	Correlations score	
	R	R−3W+40
Speed of Identification, CP610A	0.31	0.04
Spatial Orientation I, CP501B	.37	−.02
Mechanical Principles, CI903A	.07	.25
Reading Comprehension, CI614G	.20	.23
Numerical Operations (F), CI702B	.33	−.02
Numerical Operations (B), CI702B	.27	−.01
Mathematics B, CI206C	.25	.23
General Information (Nav), CE505D	.30	.15
General Information (Pilot), CE505D	.22	.10
Complex Coordination, CM701A	.27	.20

The correlations for Speed of Identification and Spatial Orientation I show conclusively that the rights score would have a significant loading on the perceptual-speed factor, whereas the error-weighted score would not.[4] The large discrepancy between the correlations with Mechanical Principles and the small difference yielded by Reading Comprehension, indicate, on the other hand, that the error-weighted score has a visualization loading, while the rights score does not. Similarly, the rights score involves some numerical variance. There is also some indication of higher verbal and spatial-relations loadings for the rights than for the error-weighted score.

While results showing such extreme differences may be rare, there is no doubt that rights and wrongs in all but power tests deserve separate consideration and study.

Evaluation.—Factor analyses of this test indicate that it is not a particularly good measure of any one function. The loadings on the various factors show that only 35 percent of the total variance of this test has been accounted for by common factors. Of this, the visualization factor accounts for 14 percent and the length-estimation factor 9 percent. The remaining 12 percent is accounted for by other factors on which the loadings are quite low. Other tests are available that have higher loadings on the visualization and length-estimation factors.

The estimated pilot validity, computed from known factor validities, (0.22), is approximately equal to the average empirical coefficient (0.20). This indicates that all the factors in this test that are valid for pilot training have been accounted for. Since only about one-half of the nonerror variance has been accounted for, it is clear that additional factors may yet be brought out by further analysis of this test. If more complete analysis uncovers another factor, or factors, with significant loadings, further refinements may be worth while, particularly in view of the fairly high navigator validity.

Path Length, CP628[5]

Just as Map Distance, CP626, was constructed to provide a test comparable to Nearest Point, CP607, but with greater face validity and reliability, so Path Length, CP628, was constructed to provide a modified form of Shortest Path, CP608. The test was not included in any classification battery, but was used for experimental analysis only.

Description.—Portions of maps, presented in halftones, are shown, which include a reference point indicated by a dot with a circle around it and three other points marked A, B, and C. Each of the three points is connected to the reference point by a heavy black line, which may follow the wandering course of a river, a path, or a road. The task of the

[4] This interpretation, and those to follow, will become clearer upon the reader's familiarization with the contents of ch. 28, A Factorial Picture of Tests and Criteria.

[5] Developed at Headquarters, AAF Training Command. Chief contributors: Maj. S. Rains Wallace and Staff of the Perceptual Research Unit.

examinee is to select the lettered point that is connected to the reference point by the shortest path.

(1) *Internal characteristics.*—The test consists of 34 items, the first 2 of which are used as practice problems. The maps vary from item to item in size and complexity. A sample problem is shown in figure 18.5.

FIGURE 18.5
SAMPLE MAP OF PATH LENGTH, CP628

(2) *Administration.*—The examinees are told that this is a test of their ability to estimate lengths of various paths on a map. It is strongly emphasized that they are to compare lengths of the paths indicated and not straight-line distances between the points. The total testing time is approximately 8 minutes, with 5 minutes allowed for the selection and marking of answers.

(3) *Scoring.*—The scoring formula is R—W/2.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 18.12.

TABLE 18.12.—*Distribution constants for Path Length, CP628B*

Group	N	M	SD
Navigators[1]	392	19.6	4.2
Pilots[2]	176	19.4	4.8
Do[3]	112	17.3	5.6

[1] In Hondo classes 42-11 to 42-16. Tested at Psychological Research Unit No. 2.
[2] In class 43K. Tested at Psychological Research Unit No. 3.
[3] In classes 44B and 44C. Tested at Psychological Research Unit No. 3.

(2) *Reliability coefficient.*—The test was administered in two separately timed halves. An estimated reliability coefficient of 0.25, corrected for length, was obtained. This figure is based on a sample of 410 unclassified aviation students (testing dates and units unidentified).

(3) *Factorial composition.*—The largest loadings are in the length-estimation (0.25), spatial-relations (0.23), general-reasoning (0.21), and visualization (0.19) factors. Its communality (0.27) almost coincides with its reliability. For a fuller picture of the factorial composition of this test, see appendix B.

(4) *Test validity.*—A sample of 550 pilots, in classes 43K, 44A, 44B, and 44C, tested at Psychological Research Unit No. 3, yielded a biserial correlation of 0.22, corrected for restriction of range, between performance in this test and the graduation-elimination criterion in primary training. The mean score for graduates was 18.12, for eliminees 16.54, and the standard deviation for both combined was 5.00. Of the sample, 74 percent was graduates. The standard deviation assumed for the unrestricted pilot stanine distribution was 2.00.

Evaluation.—Factor analysis of this test indicates that the nonerror variance has been completely accounted for by the common factors. Four percent of the total variance is accounted for by the general-reasoning factor, 4 percent by the visualization factor, and 5 percent each by the spatial-relations and the length-estimation factors. The remaining 9 percent of the total variance accounted for lies in low loadings on other factors.

Although the validity of this test for pilots is satisfactory, its very low reliability (0.25) and its factorial complexity make this test unsatisfactory.

Estimation of Length, CP631A [8]

This test was also designed to measure ability to make rapid judgments of line length, but with much higher reliability than the Shorter Line Test, CP606.

Description. (1) *Internal characteristics.*—Near the center of each page are shown five bars of constant width (1.5 mm.) and of standard lengths, arranged in order from A to E. These standards vary in length from 1.5 to 2.0 centimeters, in steps of one-tenth of a centimeter. In the first part of the test, the examinee is asked to match the length of each of a number of bars of varying length with one of these standard lengths. Each item consists of a single bar which is exactly the same length as one of the five standards printed near the middle of the page.

In the second part of the test each item consists of a bar which is exactly double the length of one of the five standard lengths. The examinee is asked to judge which of the standard lengths has been doubled in each item.

The standards appear vertically on the page, but the variable test items are placed at all angles. There are 75 items in each part of the test.

(2) *Administration.*—Full directions with one sample item precede the test proper. The total testing time is approximately 12 minutes, with 4 minutes allowed for part I and 5 minutes for part II.

[8] Developed at Psychological Research Unit No. 3. Chief contributors: Cpl. Albert A. Canfield Jr. and Lt. Robert M. Gagné.

(3) *Scoring.*—In experimental work with this test rights and wrongs were scored separately.

Statistical results.—The data given below are for examinees tested at Psychological Research Unit No. 3, unless noted to the contrary.

(1) *Distribution statistics.*—Distribution statistics are presented in table 18.13. (See also table 3.1.)

TABLE 18.13.—*Distribution constants for Estimation of Length, CP631A, based upon a sample of 771 pilots*[1]

Score	Part	M	SD
Rights	I	23.3	8.0
Wrongs	I	18.4	8.2
Rights	II	22.1	7.4
Wrongs	II	33.5	11.0
Rights	I and II	45.4	13.0
Wrongs	I and II	51.9	16.6

[1] In class 44I.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.18, a standard deviation of the phi distribution of 0.20, and a range of values from −0.15 to 0.80. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 750 unclassified aviation students tested in May 1944. Since the test is highly speeded, the low mean value is not unexpected.

(3) *Reliability coefficient.*—Minimum estimates of reliability were secured by correlating the not-quite-comparable parts I and II and correcting for length. The data are shown in table 18.14.

TABLE 18.14.—*Reliability coefficients estimated by correlating part I and part II for Estimation of Length, CP631A*

Group	Score	N	r'_{II}	r_{II}
Unclassified Aviation Students[1]	Rights	355	[2]0.41	.57
Do[1]	....do....	238	[3].40	.58
Unclassified Aviation Students and Airplane Mechanics[1]	....do....	439	[2].40	.57
Do[1]	....do....	425	[3].43	.60
Classified Pilots[4]	Rights	586	[3].48	.65
Do[4]	Wrongs	586	[3].56	.72

[1] Tested at Medical and Psychological Examining Units Nos. 6 and 8 in April 1945.
[2] Part II administered approximately 4 hours after part I.
[3] Part II administered immediately after part I.
[4] In class 44I.

(4) *Correlations between rights and wrongs.*—Extremely interesting correlations between rights and wrongs were reported for this test showing a positive association between correct and incorrect responses. The data are presented in table 18.15.

TABLE 18.15.—*Correlation coefficients between rights and wrongs for Estimation of Length, CP631A*

Group	N	....	r
Pilots in preflight training	[1]431	Part I rights vs. part I wrongs	0.10
		Part I rights vs. part II wrongs	.32
		Part II rights vs. part I wrongs	.24
		Part II rights vs. part II wrongs ..	.30
Unclassified aviation students	[2]500	Part I rights vs. part I wrongs	.21
		Part II rights vs. part II wrongs ..	.32

[1] Tested in March 1944.
[2] Tested in May 1944.

(5) *Difficulty.*—Based upon item analysis of the responses of 750 unclassified aviation students tested in May 1944, the test yielded a mean proportion of correct responses of 0.36, corrected for chance, with a range from 0.00 to 0.95, and a standard deviation of 0.21.

(6) *Test validity.*—Validation results based on two samples are given in table 18.16.

TABLE 18.16.—*Validity data for Estimation of Length, CP631A, based upon the criterion of pilot primary graduation-elimination*

Score	N_t	p_g	M_g	M_e	SD_t	r_{bis}	${}_c r_{bis}$[1]
Part I rights	[2]771	0.85	23.56	21.62	7.95	0.13	0.19
Part II rights	[2]771	.85	22.39	20.55	7.37	.14	.18
Part I wrongs	[2]771	.85	18.29	18.84	8.34	−.04	−.06
Part II wrongs	[2]771	.85	33.72	32.55	10.95	.06	.06
Part I rights	[3]431	.76	25.02	23.87	7.66	.09	.13
Part II rights	[3]431	.76	21.24	20.62	6.43	.06	.09
Part I wrongs	[3]431	.76	20.41	20.51	8.99	−.01	−.01
Part II wrongs	[3]431	.76	32.79	33.08	10.62	−.02	.00
Total rights	[3]431	.76	46.26	44.49	12.14	.09	.13
Total wrongs	[3]431	.76	53.20	53.59	15.78	−.02	−.01

[1] Corrected to an unrestricted stanine standard deviation of 2.00.
[2] In class 44I. Tested March 1944.
[3] In class 44I. Independent of sample of 771.

(7) *Item validity.*—Validation of items revealed a mean phi of 0.03 based upon the responses of 400 graduates and 80 eliminees from training in class 44I. The range was from −0.11 to +0.33, and the standard deviation was 0.08.

Evaluation.—Estimation of Length, CP631A, seems to have low to moderate pilot validity and satisfactory reliability. The test, unfortunately, has not been submitted to factor analysis, but it promises to have a substantial loading on the length-estimation factor, which is known to be valid for pilots.

The fact that the standard lengths are placed vertically on the page, while the variable lengths are at all angles, introduces the vertical-horizontal illusion into the test. This is unfortunate; future refinements of the test should be designed to investigate its effect.

The data showing positive correlations between the number of correct and of incorrect responses again directs attention to the important problem of the roles of rights and wrongs in a test. No a priori scoring formula would do justice to this peculiarity in this test. Apparently, examinees who get a high rights score do so by working rapidly; in so doing, they also make many mistakes. Those who get low rights scores go slowly and cautiously, and so get low wrongs scores also. It may be necessary to control working time by tachistoscopic exposures in order to assure maximum univocal meaning of scores in this test and in other perceptual tests of its type. In lieu of this, an empirically derived scoring formula is required.

Distance Estimation, CP212A [7]

This test is designed to measure the ability to make spatial discriminations based on the perception of distance. Other tests for the perception of extended distances or depth commonly involve the presentation of stimulus objects located at 20 feet or less from the individual being tested. The air-crew member, however, has to deal with distances in the neighborhood of hundreds of yards, for example, in landings and take-offs.

It is known that the cues that chiefly determine the perception of short distances such as those represented in other depth-perception tests, namely, the binocular cues of retinal disparity and convergence, and the monocular cue of accommodation, diminish in importance with increasing distance and finally disappear at the greater distances. It may be argued, therefore, that other tests do not measure the function of distance perception as it is defined in the pilot's task.

Distance Estimation, CP212A, is designed to involve the perception of extended distances, and the test attempts to provide the visual stimuli for the perception of such distances by means of photographs.

The informed reader will note the resemblance of this test to laboratory investigations into size constancy phenomena and may question that it is primarily a test of distance perception. The rationale of the test, as outlined by those who developed it, is as follows: the ability to discriminate size when distances are not involved is relatively accurate. On the other hand, when differences of distance are involved, the accuracy of size perception is much impaired. It is argued, therefore, that errors in the estimation of size seen at a distance are attributable more to inaccuracy in distance perception than to size perception per se.

Description.—The test consists of 2 sets of 20 glossy 9-inch x 12-inch photographs, making 40 items. In the foreground of each are 15 vertical white stakes arranged in ascending height from left to right with a wide space in the center. The stakes actually vary from 27 to 83 inches in height, differing from each other by 4 inches. They were photographed at a distance of 14 yards from the camera. In the space between the

[7] Originated at the Perceptual Research Unit, Headquarters, AAF Training Command, and developed at the Psychological Test Film Unit. Further discussion on both the predecessors and successors to this test may be found in another report (report 7) of this series.

standard stakes, and at a much greater distance from the camera, is another stake which may be one of four heights: 63, 67, 71, or 75 inches. This stake may be presented at one of 5 distances: 28, 56, 112, 224, or 448 yards. All stakes vary in width between 2 and 4 inches. The task of the examinee is to match the far, or distant, stake with one of the standards presented in the foreground. A sample photograph is presented in figure 18.6.

(1) *Administration.*—Full, detailed instructions are given with this test, calling particular attention to the fact that the stakes vary in width. Eighteen minutes are allowed for administration.

(2) *Scoring.*—The rights (R), "one-step" wrongs (W_1), and "two-step" wrongs (W_2) are scored and used in the formula: $3R+2W_1+W_2$. This formula is not based upon statistical analysis, but it is designed to credit incorrect responses in proportion to the degree of correctness they represent.

Statistical results.—The available data are for examinees tested at the Psychological Test Film Unit.

(1) *Distribution constants.*—Distributions obtained on this test are given in table 18.17.

TABLE 18.17.—*Distribution constants for Distance Estimation Test, CP212A, based upon a sample of 50 returnee air-crew members*

Scoring formula	Mean	SD
Rights	6.4	3.4
$3R+2W_1+W_2$	49.3	16.4

(2) *Reliability coefficient.*—One sample yielded the estimates of reliability given in table 18.18.

TABLE 18.18.—*Reliability coefficients (alternate forms) for Distance Estimation Test, CP212A, based upon a sample of 50 returnee air-crew members*

Scoring formula	r'_{tt}	r_{tt}
Rights	0.40	0.57
$3R+2W_1+W_2$	.66	.79

(3) *Other data.*—An experiment was performed to determine the degree of correspondence between judgments made using the photographs, and judgments obtained in the actual field situation represented by the photographs. A comparison of the judgments of 13 examinees (enlisted men and 1 civilian) in the 2 situations reveals a high degree of relationship between them. Rank-order correlations were obtained of the judgments made in the field situation and its photographic representation at each of five distances on each of four heights of test object. The median coefficient is 0.80. This is fairly good evidence that genuine distance perception can be measured with photographic representation of the distance.

Evaluation.—This test has a certain obvious value in that it attempts to measure distances that are truly representative of those that must be estimated by air crew. Although the evidence of correspondence between judgment of photographically represented distance, and distance judged under field conditions is little more than indicative, it is certainly sufficient to warrant further development of the test.

ANGULAR JUDGMENT TESTS

The three tests described in this section all attempt to assess the ability of the examinee to estimate angular magnitudes, an activity involved in the duties of all air-crew members.

Angular Judgment, CP217A [*]

This test was developed as a measure of the ability to estimate angular magnitudes.

Description.—The test consists of 45 items, in each of which a drawn angle is presented, followed by 5 possible numerical answers, one of which is correct. The angles range from 15° to 330°, and they are represented by line drawings in which is clearly indicated, by means of an arrow, the angle to be considered.

(1) *Internal characteristics.*—The directions present two sample problems. Six illustrative angles with correct answers are given along with the directions, in order to provide each examinee with some frame of reference. These illustrations, however, are not referred to during the test. The test is divided into two parts, with 23 and 22 items respectively.

(2) *Administration.*—The test was administered with various time limits, ranging from 4 to 10 minutes.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The available data all are based upon examinees tested at Psychological Research Unit No. 2.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 18.19.

TABLE 18.19.—*Distribution constants for Angular Judgment, CP217A, based upon samples of pilots in elementary training*[1]

N	Class	M	SD
281	42I; 42J	16.5	5.6
172	......[2]	17.5	6.6
208	44E; 44F	21.0	6.4

[1] Allowing 5 minutes per part.
[2] Class unidentified. Examinees were tested in October 1943.

(2) *Reliability coefficient.*—By the odd-even method, an estimated reliability coefficient of 0.87, corrected for length, was obtained. This figure is based on a sample of 734 unclassified aviation students.

[*] Developed at Psychological Research Unit No. 2. Chief contributors: Capt. John T. Dailey, Capt. Glen Finch, Tech./Sgt. Paul W. McReynolds, and Tech./Sgt. Benjamin Shimberg.

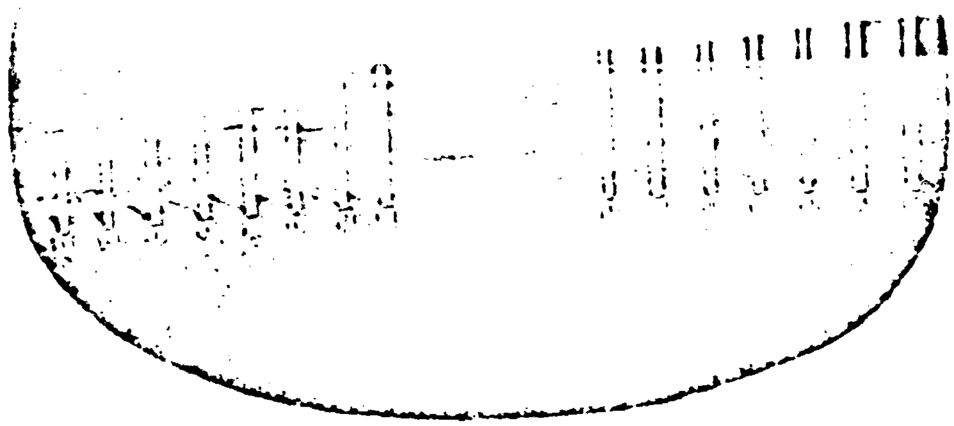


FIGURE 18.6
SAMPLE PHOTOGRAPH OF DISTANCE
ESTIMATION, CP212A

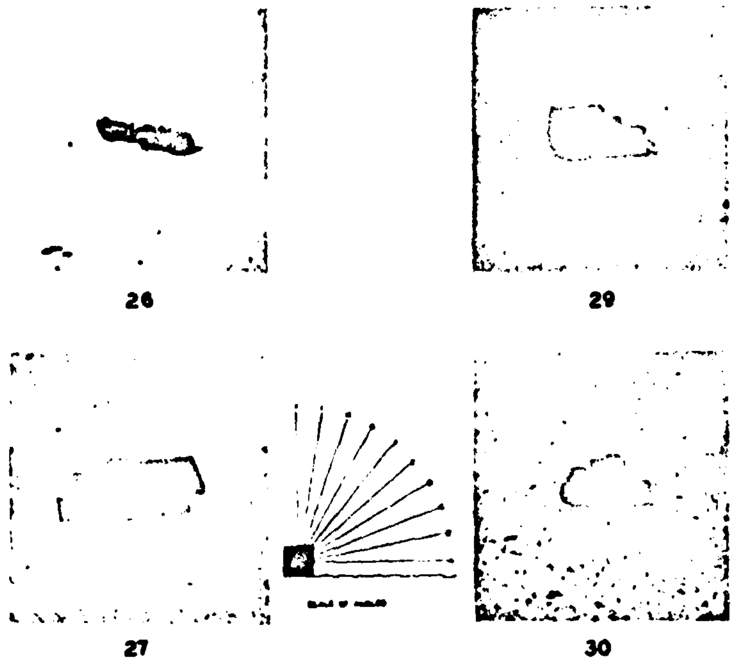


FIGURE 18.7
SAMPLE ITEMS AND SCALE OF ANGLES
OF ANGLE ESTIMATION, CP218A

(3) *Test validity.*—A sample of 972 pilots in classes 44F and 44I yielded a biserial correlation of 0.20, corrected for restriction of range, between performance in this test and the graduation-elimination criterion in primary training. The mean score for the graduates was 18.11, for eliminees 16.73, and the standard deviation for both combined was 6.35. Of this sample 81 percent were graduates, and the standard deviation assumed for the unrestricted pilot stanine distribution was 2.00.

Evaluation.—This test is moderately good as a predictive device for pilot training (validity coefficient of 0.20). It also has a satisfactory reliability (0.87). Considering these facts and, in addition, the simplicity of design and administration, further analysis and development of this instrument seems worth while. It is possible that this test, and the tests Angle Estimation and Landing Judgment, to be described next, will define a new factor.

Angle Estimation, CP218A [9]

This test is designed to measure the ability to estimate the angle at which an object on the ground is viewed from various points above it in the air. From various sources, such as job analyses and subjective judgments, there are indications that the ability to estimate such angles may be very important in landing a plane. The pilot who more readily and accurately estimates the angle of his approach to the landing strip should, other things being equal, make the better landing.

Description.—The test is composed of photographs of models of military vehicles, such as trucks, jeeps, and tanks, taken at angles ranging from 0° to 90° between the camera axis and the plane of the ground.

The angles at which the photographs were taken are in multiples of 10 degrees. The task of the examinees is to indicate the angle at which the photograph was taken. On each page of the test there is a scale of angles such as that shown in figure 18.7. This scale shows 10 different angles, ranging from 0° to 90°, and each labeled with a letter symbol. The examinees indicate the angle of the photograph by using the appropriate symbol.

(1) *Internal characteristics.*—There are 48 items in the test, including 2 unscored practice problems. The test is divided into two comparable parts. The photographs have a common background of a sandy surface, and none has any horizon line, showing no sky.

(2) *Administration.*—The examinees are informed that this is a test of their ability to estimate the angle between the line of sight and the surface of the ground. The scale of angles is explained as follows:

All photographs in the test have been taken from 1 of the 10 angles (of the scale of angles) above, lettered A through J. The angles are at intervals of 10 degrees. It will be your task to study the picture in each item and estimate the angle from which it was taken.

[9] Developed at Psychological Research Unit No. 3. Chief contributors: Sgt. Roy C. Anderson, Staff/Sgt. Benjamin Fruchter, and Lt. John W. Howe Jr.

703320—47—31

The total testing time is approximately 15 minutes, with 5 minutes allowed for each part.

(3) *Scoring.*—In the scoring of this test, correct answers originally received +1, answers that were "one-step" wrong received 0, and those "more-than-one-step" wrong received −1. A study was made to determine the scoring formula which would maximize reliability, utilizing a sample of 736 unclassified aviation students. This formula was found to be R−2W, "one-step" wrongs being eliminated from consideration. With a constant added to eliminate negative scores, this became R−2W+60.

Statistical results.—The data given below all are for examinees tested at Psychological Research Unit No. 2 in February and March 1945.

(1) *Distribution statistics.*—Scoring rights only, a mean score of 17.3 and a standard deviation of 4.8 was found for a sample of 736 unclassified aviation students.

(2) *Internal consistency.*—The degree of homogeneity of the items of the test is indicated by a mean internal-consistency phi of 0.26, a standard deviation of the phi distribution of 0.11, and a range of values from 0.00 to 0.49. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score (rights only) of a group of 736 unclassified aviation students.

(3) *Reliability coefficient.*—The reliabilities were estimated separately for rights, wrongs (more-than-one-step), and wrongs (one-step). A sample of 736 unclassified aviation students was utilized. Kuder-Richardson data appear in table 18.20, and intercorrelations among the various scores are presented in table 18.21.

TABLE 18.20.—*Kuder-Richardson (formula no. 21) reliabilities of three scores for Angle Estimation, CP218A*

Score	M	SD	r_{11}
Total rights	17.3	4.8	0.54
Total wrongs (more-than-one-step)	9.6	5.1	.72
Total wrongs (one-step)	20.9	3.7	.14

TABLE 18.21.—*Product-moment correlations among the various scores for Angle Estimation, CP218A*

Score	1	2	3	4	5	6
1 Rights, part I	...	0.39	−0.62	−0.45	−0.38	0.10
2 Rights, part II	0.39	...	−.41	−.61	.03	−.38
3 More-than-one-step wrongs, part I	−.62	−.41	...	.5[illegible]	−.36	−.16
4 More-than-one-step wrongs, part II	−.45	−.61	.54	...	−.07	−.46
5 One step wrongs, part I	−.38	.03	−.36	−.07	...	.03
6 One-step wrongs, part II	.10	−.38	−.16	−.46	[illegible]	...

It can be seen that one-step wrongs are highly unreliable.

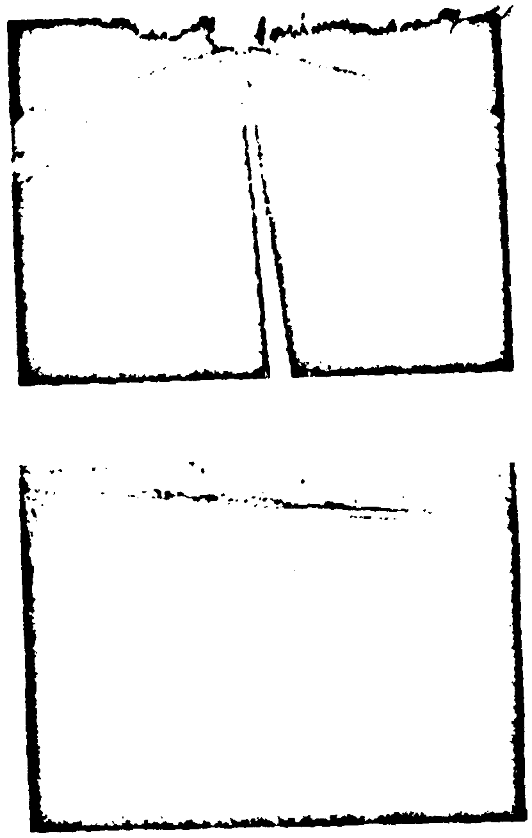


FIGURE 18.8
SAMPLE ITEMS OF LANDING JUDGMENT.
CP503B

For this sample, the maximum-reliability scoring formula would yield a part I-part II correlation of 0.58. Correcting for length, this becomes 0.73.

(4) *Difficulty.*—Based upon item analysis of the responses of 750 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.23, corrected for chance, with a range from 0.05 to 0.88 and a standard deviation of 0.19.

Evaluation.—This test, with Angular Judgment, CP217, may measure a new factor. The difficulty level, however, is unsatisfactory. The data for the reliabilities of rights and the two wrongs scores suggest that the differences utilized in the photographs are not gross enough to provide the most reliable score. Some preliminary psychophysical determinations are strongly indicated.

Landing Judgment, CP505B [10]

This test was designed to measure height judgment. Analysis of the act of landing an airplane showed that students lacking in height judgment are bad risks.

Description.—Specifically, the test requires the examinee to learn, remember, and later select the point at which to break the glide in landing. It consists of photographs simulating the pilot's view of a landing strip, road, or field as he comes in for a landing. Figure 18.8 is an example of the test items.

(1) *Internal characteristics.*—A period of training precedes the test proper. At the beginning of the test, the examinees are shown three views of a landing strip, each taken at a different height. One, they are told, shows the correct height (15 feet); another is too low (5 to 10 feet); and the last too high (20 to 25 feet). The examinees are to judge each photograph, marking A if it is the correct height for leveling off; B if it is too low; and C if it is too high. The photographs are printed each showing about the same amount of sky above the horizon; in this way, one cue was eliminated. The cues contributing to correct judgment are linear perspective and surface texture.

The examinees then study another series of three photographs, this time of a grassy field. After administrative instructions, they are allowed 2 minutes to restudy these first six photographs.

Following this reinforcement, the examinees answer four sample problems. After putting down their answers, the examinees are told the correct answers.

The test is divided into 2 parts of 26 items each.

(2) *Administration.*—The examinees are told that:

This is a test of your ability to learn and remember what the ground looks like from the correct height for leveling off during the final part of the glide. For this test, 15 feet above the ground has been selected arbitrarily as the correct height for

[10] Developed at Psychological Research Unit No. 3. Chief contributors: Pvt. Charles K. Ferguson, Cpl. Harold H. Kelley, Staff/Sgt. Wayne S. Zimmerman.

leveling off . . . The photographs have been taken in such a way that you will not be able to determine the height by the amount of sky in the picture or the place where roads or fields are cut off at the edges of the picture. Instead you must judge or "feel" the height from the picture as a whole.

The examinees are allowed to study the picture giving the correct heights and heights too low and too high during the reading of the directions. After the four sample problems are worked, 1½ minutes are allowed for the reexamination of these problems before the test begins. The total testing time is approximately 12 minutes, with 3 minutes allowed for the first part, and 2 minutes and 45 seconds for the second part.

(3) *Scoring.*—The scoring formula is $R-W/2$.

Statistical results.—No statistics are available for this test.

Evaluation.—This test is obviously an attempt to duplicate a part of the landing situation as nearly as possible using paper and pencil. It seems probable that it will have loadings on several factors. It may, however, help to define a new factor held in common with the Angle Estimation and Angular Judgment tests. The approach underlying the construction of this test, that is, constructing tests that approximate a work sample, largely ignores the individual basic factors involved in the activity. This differs from the approach stressed in the latter part of the classification program—the construction of pure tests.

JUDGMENT OF PROPORTIONS TEST

Only one test in this area was constructed.

Judgment of Proportions, CP206B [11]

This test is designed to measure the ability to recognize accurately the correct proportions of familiar objects. The rationale underlying this test is that accuracy of object perception is important for air-crew personnel, and that successful recognition of the correct proportions of familiar objects reflects previous interest, accuracy of perception, and good visual memory for seen objects and for spatial dimensions in general.

Description.—The examinee is asked to select, from five simple outline diagrams, the one that has the correct proportions of a given familiar object, such as a standard building brick, a one-dollar bill, a pack of standard-size cigarettes. Despite the fact that all items in the test actually have three dimensions, only two are depicted, and these are specified. Two items are presented in figure 18.9. There are 30 items in the test.

[11] Developed at Headquarters, AAF Training Command. Chief contributors: Capt. Richard H. Henneman and Staff of the Perceptual Research Unit.

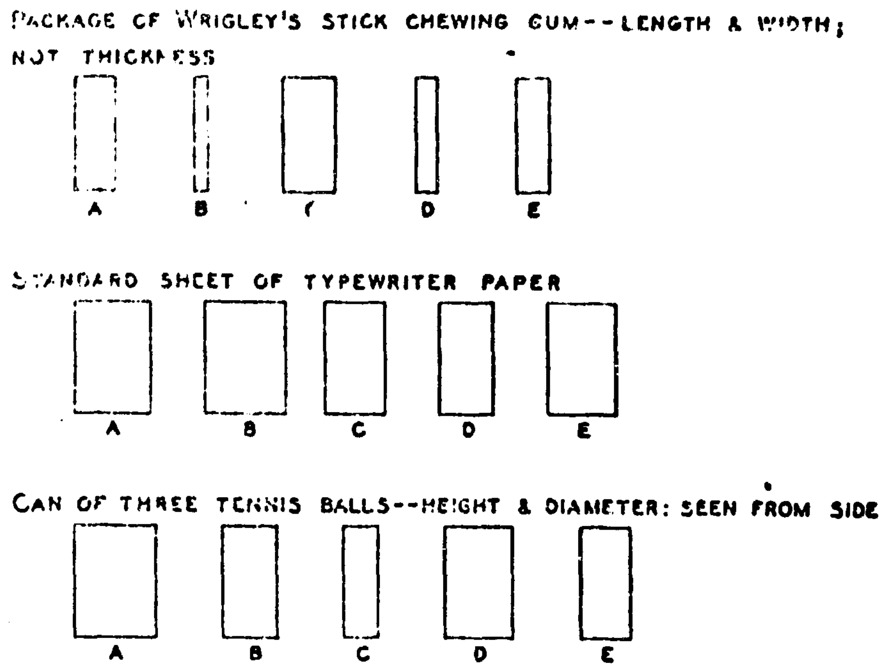


FIGURE 18.9
SAMPLE ITEMS OF JUDGMENT OF PROPORTIONS, CP206B

(1) *Administration.*—The examinees are told that this is a test of ability to judge the shape of familiar objects. Particular attention is called to the directions concerning the dimensions of the objects that are to be considered. Testing time for the 30 items is 8 minutes.

(2) *Scoring.*—The scoring formula is R−W/4.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 18.22. The distribution curves are approximately symmetrical and somewhat flatter than normal.

TABLE 18.22.—*Distribution constants for Judgment of Proportions, CP206B, for samples of unclassified aviation students*

N	M	SD
1,098	10.9	4.4
392	12.0	4.6

(2) *Reliability coefficient.*—By the odd-even method, an estimated reliability coefficient of 0.52, corrected for length, was obtained. This figure is based on a sample of 1,098 unclassified aviation students.

(3) *Factorial composition.*—The most significant loadings are in the planning (0.30), visualization (0.29), verbal (0.22), and perceptual-speed (0.22) factors. The communality is 0.35. For a fuller picture of the factorial composition of this test, see appendix B.

(4) *Test validity.*—Validation results based on several samples are given in table 18.23.

TABLE 18.23.—*Validation data for Judgment of Proportions, CP206B, for the graduation-elimination criterion*

Group	Class	N_t	p_s	M_s	M_e	SD_t	r_{bis}
Pilots in primary training	43H; 43I	[1]380	0.60	11.51	10.95	4.38	0.08
Do	43K	[2]178	.87	12.93	11.58	4.47	.16
Flexible gunnery students ...	......	[3]327	.93	13.51	11.73	1.78	.24

[1] Tested at Psychological Research Unit No. 1.
[2] Tested at Psychological Research Unit No. 3.
[3] Unidentified.

Evaluation.—This test was included in two factor analyses. The weighted average of the factor loadings shows that 35 percent of the total variance has been accounted for by the common factors. Of this, the verbal and perceptual-speed factors each contribute 5 percent, the visualization factor 8 percent, and the planning factor 9 percent. The remaining 8 percent is accounted for by other factors on which the loadings are quite low. Since the test was designed to measure visual imagery, among other things, it is moderately successful, for the visualization factor accounts for about a quarter of the common-factor variance. In terms of the total variance, however, the visualization portion is quite small. In addition to this fact, the test has a very low reliability (0.52), and the problem of presenting objects equally familiar to all examinees was not solved satisfactorily.

The elementary pilot validity coefficient is quite low (weighted average approximately 0.08). An estimated validity coefficient, computed from the valid factor loadings (see ch. 28) is 0.18, which suggests a sampling error in the obtained validity. At least one can say that all the factors in the Judgment of Proportions test that are valid for pilot training have been fully accounted for.

This test utilizes subject matter different from any other, but there exists, in other tests, better measures of the valid factors included in Judgment of Proportions, and these tests have a considerably higher reliability. For these reasons, further work with the test was not considered worth while.

SUMMARY AND EVALUATION

Under the category of size and distance perception, seven printed tests involving li[illegible] nonlinear extents, three involving angular magnitudes, and [illegible]lving perception of proportion, were developed and studied.

Little evaluation can be given at present of the three tests of angular estimation, since factorial data are not available, and validity data are available for but one of the tests. There is an interesting possibility, however, that they may define a new factor, perhaps valid for one or more of the air-crew positions. The available evidence suggests that Angle Estimation be revised, utilizing slightly grosser stimulus differences.

The Judgment of Proportions test seems to be of little value. It had little or no validity to show for the pilot criterion, though from its factor content one would have expected more. Its reliability is highly unsatisfactory; and it is factorially complex, with no very high loading on any factor yet established. There is an indication that some of the variance of the test has not been accounted for, but it holds no promise for pilot validity.

The other tests discussed in this chapter, perhaps with the exception of Distance Estimation, CP212A, may be considered as a group. It should be noted, first, that none of the tests for which data are available is a very satisfactory measure of the length-estimation factor. Of the tests considered in this chapter, the highest loading secured on length-estimation was 0.46, for the Shortest Path test.[11] This test, however, is contaminated with loadings in the perceptual-speed, visualization, and spatial-relations factors. Shorter Line, with a loading of 0.44, is much more pure. Increasing its reliability should make this test quite satisfactory. This was attempted in the Estimation of Length test, for which factorial data are not yet available.

Comparing the factor patterns of Shorter Line, Nearest Point, Shortest Path, Map Distance, and Path Length, we find very strong evidence that the best length-estimation test requires a simple comparison of two linear lengths presented as continuous, uninterrupted extents, regularly placed in one orientation, either horizontal or vertical. The introduction of extents bounded by points, of curved or irregular lines, of distracting background, of irregular placement (introducing the vertical-horizontal illusion)—all seem to bring in other factors, particularly visualization. The results for one test (Pattern Assembly; see ch. 17), however, are against these conclusions. A decision must await further factorial evidence.

Three other points are worth mentioning. First, the fact that all the tests considered in this chapter deal with limited sizes and distances raises the question of the generality of the length-estimation factor. Will tests of the ability to perceive correctly much greater distances involve the same factor? Evidence from experimental psychology suggests an affirmative answer to the question.

Secondly, the analysis of two of the tests in this chapter served to call sharp attention to the potentially distinct roles of incorrect and correct responses, and to the dangers inherent in the indiscriminate use of a priori scoring formulas. The data for the Map Distance test dramatically show how the factorial composition of a test may change with varying penalties for incorrect responses. The data for the Estimation of Length

[11] Pattern Assembly, CP804A, has a loading of 0.52 on the length estimation factor. See the discussion of this test in ch. 17.

test show a well-substantiated positive correlation between numbers of correct and incorrect responses.

Thirdly, the promise of the Distance Estimation test, utilizing photographic representation of extended distances, is most encouraging. If the correlation of distance estimation in three-dimenional space with estimation in space represented in two dimensions is well established in a large-scale study, and if the correlation is sufficiently high, the potential usefulness of printed tests will be greatly extended.

CHAPTER NINETEEN

Spatial Tests[1]

RATIONALE FOR SPATIAL TESTS

The Concept of Spatial Ability

The concept of a measurable spatial ability was brought to attention primarily as a result of factor analyses of intellectual abilities. Corroborating the earlier work of Kelley (1), Thurstone listed a factor S, described as "facility in spatial and visual imagery" among his seven primary mental abilities (2).

Unlike other chapter designations in this volume the category "spatial tests" is not an old and familiar term in differential psychology. In fact, the concept is still being formed. The concept of spatial ability, like other concepts born of factor analysis, was developed inductively. In the inductive approach, statistical evidence is the basis for deriving a new psychological concept. For many of the older psychological concepts almost the reverse has been true. A variable in which individuals differ would be assumed, then tests would be built to measure it, and statistical results would be compiled to describe individuals and populations with respect to it.

Factor Studies in Aviation Psychology

In the work of the AAF Psychological Research Units, an effort to study the nature of spatial abilities was first stimulated by the appearance of a new and unidentified factor in an analysis of a battery of tests designed to measure foresight-and-planning ability (see ch. 9). For want of a better term, the factor was labeled the "nonmotor" or, sometimes, the "intellectual" component of Complex Coordination, since its highest loading was in that test.[2] It was recognized that this label left much to be desired, but the diverse nature and limited number of tests permitted no more positive designation. It was suggested as one hypothesis, however, that this might be a spatial factor, since the tests with substantial loadings called for the correlation of spatial ar. ngements in stimulus and response.

This new find' 'g was of particular interest, because the pilot validity of this test was very high; yet its predictive value was not fully accounted for either by the factors already identified, including mechanical

[1] Written by Lt. John W. Howe Jr. and Staff/Sgt. Wayne S. Zimmerman.
[2] See page 122 for a brief description of this test, and Report No. 4 for a complete description.

experience and perceptual speed, or by assumed motor-coordination factors. If Complex Coordination's validity were due in part to its loading on the new factor, important implications could be drawn. It was decided that the new factor should be fully explored and identified. Since it appeared to be measurable by means of printed tests, it was felt that efforts should be made to maximize its content in such more economical tests.

A few months subsequent to the aforementioned analysis, two other analyses were completed, both of which verified the new factor. In the first, Instrument Comprehension II, CI616B, Complex Coordination and Planning Air Maneuvers appeared together; and in the second, Planning A Course, CI406AX2, Planning Air Maneuvers, and Complex Coordination defined a factor.

The new tests were scrutinized for further clues. Planning A Course and, to a lesser extent, Instrument Comprehension II, formerly had been described as tests of "integrative ability," i. e., the ability to synthesize quickly the influence of several environmental factors that bear upon the choice of a single direction of action. Accordingly, the hypothesis was entertained for a while that the factor was one of integrative ability. But later evidence weakened this belief and supported instead the earlier surmise that the factor was, in reality, one of spatial ability.

Later, Psychological Research Unit No. 3 completed analyses of two other batteries, both of which confirmed the factor. In the first of these analyses, Discrimination Reaction Time, CP611D, Complex Coordination, Two-Hand Coordination, CM101A, and Dial and Table Reading, CP622A and CP621A, appeared together with strong projections on a common reference axis. In the second analysis the following tests had substantial projections on a single axis: Complex Coordination; Flags, Figures, and Cards, CP512A; Directional Orientation, CP515; Cubes, CP512A; Table Reading, CP621A; and Dial Reading, CP622A.

Purposely included in the second correlational matrix were two of the tests that Thurstone had originally found to determine his spatial factor, namely, Cubes, and Flags, Figures, and Cards. Their emergence on the factor supported the case for the spatial hypothesis.

In attempting to define this spatial ability in more specific terms, a great deal of difficulty has been encountered. As mentioned in the introductory paragraph of this chapter, Thurstone described his spatial factor as "facility in visual and spatial imagery." He failed to isolate an additional factor that could be called visualization, although further rotations involving one of his residual axes would have revealed it. Actually the expression "facility in the use of visual imagery" better describes a factor which in this volume is labeled visualization.

The precise definition of the spatial factor must wait until more crucial results are available. It is generally agreed that the factor is a spatial one, but beyond that point there are divergences of opinion. Two hypotheses which have been proposed deserve consideration here. According to one hypothesis, it is an ability to make discriminations as to direction of motion. The term "discrimination" here does not carry the usual connotaton of perceiving small differences, for obviously the spatial distinctions called for in the tests are very gross. The decisions are frequently merely between up and down, left and right, and out and in.

Another hypothesis is that the ability is concerned with the general apprehension of spatial relations. Either stimuli or responses or both in the spatial tests are arranged in spatial patterns, and there is frequently a systematic relationship between order in the stimulus and order in the response. In such a test, therefore, the essence of the spatial factor could be (1) ability to pereceive visual-spatial arrangements, (2) ability to organize movements in spatially-determined order, or (3) ability to relate specific spatial locus or arrangement within the stimulus pattern with specific locus or arrangement within the response pattern. The second and third characteristics would apply to the psychomotor tests but not to all printed tests. The first of these three must therefore be the most significant.

The tests in this chapter are divided into two subareas. In keeping with the policy followed in general throughout this volume, the subarea designations are selected on logical evidences which follow from the surface characteristics of the tests rather than on factorial grounds. Since the tests in the first subarea tend to call for distinctions as to direction, the title selected for the group is "Directional Discrimination Tests." The second section will be entitled "Positional Discrimination Tests." As it turned out, the division of tests along these lines coincides quite well with the distinction between two space factors, S_1 and S_2, which will be described later in this chapter.

DIRECTIONAL DISCRIMINATION TESTS

The following seven tests are treated in the first subarea: Instrument Comprehension I, CI615A, B; Instrument Comprehension II, CI616A, B, C; Aerial Orientation, CP520A, B; Flight Orientation, CP528A; Stick and Rudder Orientation, CP531A; Discrimination Reaction Time (paper), CP634A; and Directional Marking, CP533A.

Instrument Comprehension, CI615B [3] and CI616B [4]

This test consists of two parts sufficiently distinct to warrant two code numbers. Each was designed to measure certain aspects of the ability to comprehend a plane's behavior on the basis of instrument readings. When the test idea was first proposed in July 1942, it was called Dial

[3] Developed at Psychological Research Unit No. 3. Chief contributors to the original form CI615A: Capt. Milton Burdman and Lt. Wilbur S. Gregory. Chief contributor to form CI615B: Lt. David H. Jenkins.

[4] Developed at Psychological Research Unit No. 3. Chief contributor to the original form CI616A: Capt. Milton Burdman. Chief contributor to form CI616B: Lt. David H. Jenkins.

Interpretation, and the following rationale was presented:

This test will measure several abilities:
(1) Ability to visualize the behavior of the plane in space.
(2) Ability to translate a verbal description of the plane's behavior into a visual image of the plane in that maneuver.
(3) Ability to interpret dial readings in terms of the plane's behavior.
(4) Skill at dial reading.
(5) Understanding of the relations between 1, 2, and 3 above.
(6) Ability to keep in mind several factors at one time.

Description.—Form CI615B-CI616B is chosen for presentation here, since it is the most representative form and the first to be included in the classification battery.

For each item in part I (CI615B), the examinee mu . refer to drawings of six instruments—altimeter, compass, air-speed indicator, artificial horizon, rate-of-climb dial, and turn-bank indicator. He must then select the correct one of five written descriptions of a plane's position in flight.

In part II (CI616B), each item presents drawings of only two instruments, compass and artificial horizon, followed by five small photographs showing an airplane in five different positions, e. g., headed north and climbing, headed northwest and diving, etc. The examinee must choose the picture which is in agreement with the two instrument readings.

Part II seems to be superior to part I on an a priori basis. It requires the examinee to relate the dial readings and plane's behavior in a much more direct and simplified manner than part I. Also, it does not involve a possible reading-speed factor and the verbal-comprehension factor which seem to be apparent in part I. The verbal nature of the items in part I is comparable to the type of directions that instructors give student pilots, however, so that part I possibly makes a contribution to the test that part II does not. Also, because part I involves several dials, it does measure ability to attend to several factors at once to a greater extent than part II does. In view of these facts, it was originally believed that the test should include both types of items until their validities had been experimentally determined. It will be seen later that Instrument Comprehension proved to be quite valuable, but not exactly for the theoretical reasons first stated. It remained for factor analysis to reveal the underlying sources from which the test derived the high validity that it achieved.

(1) *Internal characteristics.*—Part I contains 1 unrecorded and unscored practice item and 15 recorded and scored test items. Part II contains 60 scored items.

(2) *Administration.*—For part I the instructions and the single practice problem following them require about 10 minutes. Working time for part I is 12 minutes, which allows about 80 percent of the examinees to finish. The time for part II is 4 minutes for instructions and 15 min-

utes working time. Approximately 40 percent of the examinees finishes in the allotted time. Thus, in this part, a greater premium is placed upon speed.

Dials for sample A are shown in figure 19.1. The following are excerpts from the directions for part I:

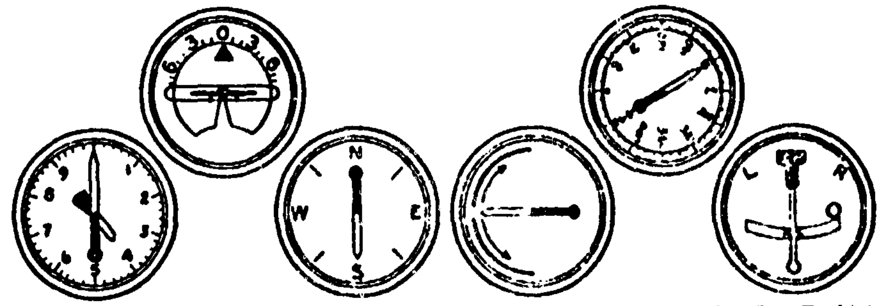


FIGURE 19.1
DIALS FOR SAMPLE PROBLEMS,
INSTRUMENT COMPREHENSION I, CI615B

This is a test of your ability to interpret dial readings on the instrument panel of an airplane.

Six of the dials commonly found in military planes are used. Dial I is * * *

Each of the six dials is then quite fully explained and illustrated. Following these explanations, the directions continue:

In each problem of part I diagrams of the six dials appear on the left page of the booklet. Opposite the dials, on the right page, are five written descriptions of the plane's behavior. You are to examine the dials on the left page, then choose the correct description from those appearing opposite the problem on the right page.

Sample Problem A

A. Flying level at 200 m. p. h. straight and unbanked, headed due south, gaining altitude at 9,800 feet.
B. Flying level at 200 m. p. h. straight and unbanked, headed due south, losing altitude at 5,000 feet.
C. Flying level at 200 m. p. h. straight and unbanked, headed due south, maintaining altitude at 4,000 feet.
D. Flying level at 200 m. p. h. straight and banked to left, headed due north, maintaining altitude at 4,000 feet.
E. Flying level at 200 m. p. h. turning properly to left, with 30° bank, maintaining altitude, headed due north at 4,000 feet.

Description C is the correct answer.

Part of the directions from part II follow, and a sample problem is shown in figure 19.2.

In each of the problems in part II you will be given a picture of a single plane in five different positions. At the left of the picture you will be shown two dials, an artificial horizon and a compass. You are to choose the position of the plane which agrees with the readings on these dials.

Now examine the dial readings at the left of the pictures in the sample problem. Then look at the five positions of the plane and select the correct position.

According to the dials, the plane is banked left, flying level, and is headed south. (B) is the correct answer, because at position (B) the plane is banked left, flying level, and is headed south.

(3) *Scoring.*—The scoring formula used for part I (CI615B) is of particular interest, because it is the only instance in any classification battery in which correct answers actually decrease the score. The formula for the score on part I was 20 − (R − W/4). The reason for giving negative weight to correct answers is that part I was found to possess a low positive validity and such high correlations with other tests that its beta weight was negative. Accordingly, in scoring, the axes were reversed, with high scores becoming low scores and vice versa. The data presented below, however, are for the formula R−W/4.

Part II was scored simply R−W/4.

Statistical results.—Since instrument comprehension was a classification-battery test, the statistical data are very complete.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 19.1. The distribution curves for part I and part II are negatively skewed and considerably flatter than normal.

TABLE 19.1.—*Distribution constants for Instrument Comprehension, CI615B and CI616B*

Form	Group	N	M	SD
CI615B	Unclassified aviation students (Pre-CTD)[1]	[2]1,920	8.3	3.4
CI616B	Unclassified aviation students (Pre-CTD)	[2]1,920	26.6	10.7
CI615B	Unclassified aviation students (Post-CTD)	[3]1,500	10.6	3.3
CI616B	Unclassified aviation students (Post-CTD)	[3]1,500	29.6	10.8
CI615B	Armament trainees[4]	[5]269	12.5	2.9
CI616B	Armament trainees[4]	[5]269	22.6	9.3

[1] CTD stands for College Training Detachment.

[2] Tested at Medical and Psychological Examining Units Nos. 4 through 10 with the November 1943 battery.

[3] Tested at Psychological Research Unit Nos. 1, 2, and 3 with the November 1943 battery.

[4] Previously eliminated from pilot training.

[5] In Lowry Field classes 34-44A and 35-44A. Sample consists of examinees tested at all units with the November 1943 battery.

(2) *Internal consistency.*—The degree of homogeneity of part II of the test is indicated by a mean internal-consistency phi of 0.38, a standard deviation of the phi distribution of 0.14, and a range of values from −0.16 to 0.68. The criterion used was the score on part II. These statistics are based upon analysis of the responses of the highest 25 percent and lowest 25 percent in total score of a group of 800 unclassified aviation students, tested in October 1943 at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—Three samples yielded the estimates of reliability given in table 19.2.

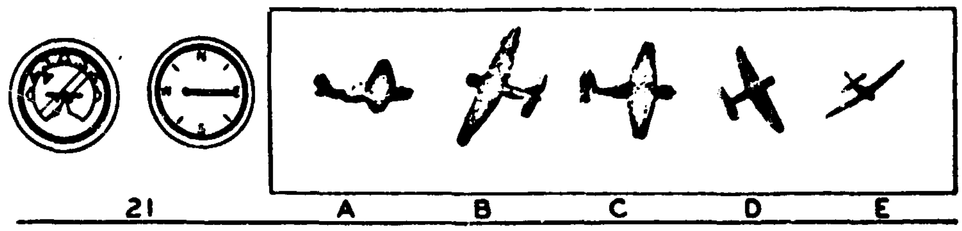


FIGURE 19.2
SAMPLE ITEM OF INSTRUMENT COMPREHENSION II,
CI616B

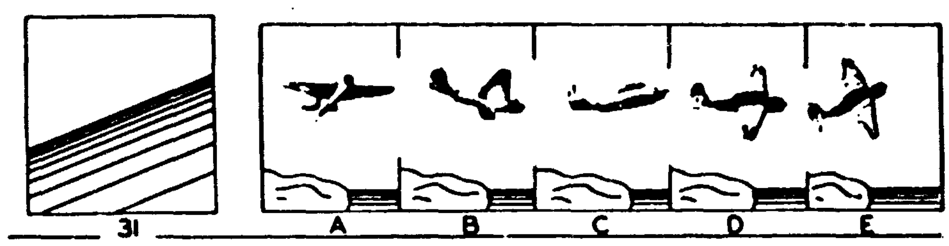


FIGURE 19.3
SAMPLE ITEM OF AERIAL ORIENTATION,
CP520A

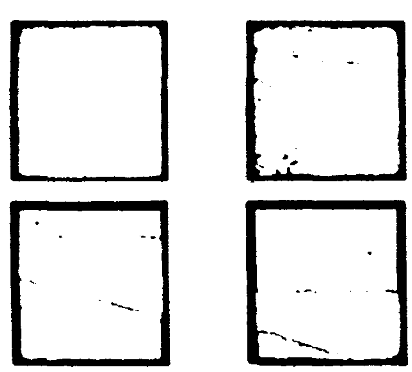


FIGURE 19.4
SAMPLE ITEMS OF FLIGHT ORIENTATION,
CP528A

TABLE 19.2.—*Estimated reliability coefficients for Instrument Comprehension based upon samples of unclassified aviation students*

Form	N	Type	r'_{tt}	r_{tt}
CI615B ..	[1]500	Odd-Even	0.68	0.81
CI616B[2] ..	[1]500	Odd-Even	.85	.92
CI615B ..	[3]1,000	Alternate-forms	.52	.68
CI616B[2] ..	[3]1,000	....do	.88	.94
CI616C ..	[4]1,115	....do	.72	.84

[1] Tested at Medical and Psychological Examining Unit No. 10 with the November 1943 battery.
[2] Note that this form is highly speeded; odd-even reliability estimates, therefore, are spuriously high.
[3] Tested at Medical and Psychological Unit No. 7 with the November 1943 battery.
[4] Tested at Medical and Psychological Examining Unit No. 8 in early 1945.

(4) *Difficulty.*—Based upon item analysis of the responses of the above-mentioned sample of 800 unclassified students, part II of the test yielded a mean proportion of correct responses of 0.68, corrected for chance, with a range from 0.24 to 0.98 and a standard deviation of 0.17.

(5) *Factorial composition.*—The most prominent loadings for part I (for the preliminary form CI615A; see below) are in the spatial-relations (0.44), reasoning II (0.34), psychomotor III (0.24), verbal (0.22), general-reasoning (0.21), and integration II (0.21) factors. The most prominent loadings for part II are in the spatial-relations (0.53), reasoning II (0.36), visualization (0.25), and verbal (0.24) factors. The communality for part I is 0.57 and for part II 0.65. For a full picture of the factorial composition of this test see Appendix B.

(6) *Test validity.*—Validation results based on several samples are given in tables 19.3 to 19.6 for form A as well as II Form B. (See below for a brief description of Form A.)

(7) *Item validity.*—Validation of items of this test disclosed the results recorded in table 19.7.

Variations.—CI615A–CI616A was the original experimental form, in which part I contained 15 items, and part II, 30 items. Part I had only low validity, while part II's validity was high enough to indicate that it should be amplified in a revision.

CI615B–CI616B was the first experimental revision, in which part II was expanded to 60 items. Part I remained unchanged.

CI616C[5] is the final modification which replaced CI615B–CI616B in the classification battery in September 1944. It consists of the 60 items of part II, with a new set of directions. Part I is entirely eliminated except for portions of its directions explaining the functions of the dials.

Evaluation.—Instrument Comprehension played a very interesting and significant role among the tests used in the psychological program. It was admitted to the classification battery in November 1943, after more than a year of development. Early data indicated that part I had a negative beta weight. Later data resulted in a zero or slightly positive weight. Thus part I was a candidate for withdrawal, as soon as it could be established that performance on part II would not be adversely affected by

[5] Developed at Psychological Research Unit No. 3. Chief contributor: Lt. David M. Jenkins.

TABLE 19.3.—*Validity data for Instrument Comprehension, CI615A[1] and CI615B, using the graduation-elimination criterion*

Group	Class	Psychological Research Unit No.	Form	N	r_t	M_g	M_e	SD_s	r_{bis}	$_cr_{bis}$
Pilots in primary training	43G, 43I	3	CI615A	666	0.80	9.2	8.3	3.5	0.15	...
Pilots in basic training	43I, 43J	3	CI615A	401	.93	8.7	6.8	3.3	.28	...
Pilots through basic training	43I, 43J	3	CI615A	533	.70	8.7	7.5	3.3	.23	[2]0.25
Pilots in primary training	44C	3	CI615B	877	.90	11.1	10.1	2.9	.18	...
D	44E	3	CI615B	585	.92	9.4	8.3	3.1	.18	[3].24
Do	44F	3	CI615B	479	.90	9.8	7.9	3.0	.33	[3].39
Do	44H	2	CI615B	504	.92	10.01	8.56	2.97	.24	[3].30
Do	44H	3	CI615B	406	.93	9.49	9.38	2.70	.16	[3].20
Do	44I	1, 2, 3	CI615A	3,146	.84	9.50	8.87	3.14	.11	[4].18
Do	44J	3	CI615B	1,676	.89	8.99	8.07	2.99	.16	[5].22
WASPS	44-W-6	..	CI615B	104	.61	10.02	7.85	3.20	.42	...

[1] Form CI615A contains the same items as CI615B. See page 483.
[2] Assumed unrestricted stanine standard deviation not reported.
[3] Assuming an unrestricted stanine standard deviation of 1.85.
[4] Assuming an unrestricted stanine standard deviation of 1.90.
[5] Assuming an unrestricted stanine standard deviation of 1.83.

TABLE 19.4.—*Validity data for Instrument Comprehension, CI615B, using various criteria*

Group	Unit	Criterion	N	r[1]
Navigator trainees	PRU No. 3	Grades in flight missions	200	0.26
Do	PRU No. 3	Grades in ground missions	200	.29
Do	PRU No. 3	Weighted average grades	200	.35
Armament trainees[2]	MPEU Nos. 1–10	Average grades	269	.23
Officer candidates	MPEU No. 5	Eighth week academic average	343	.26
Do	MPEU No. 5	Eighth week ratings	343	.14
Do	MPEU No. 5	Twelfth week ratings	343	.10

[1] Product-moment correlations.
[2] In Lowry Field classes 34–44A and 35–44A.

TABLE 19.5.—*Validity data for Instrument Comprehension, CI616A*[1] *and CI616B, using the graduation-elimination criterion*

Group	Class	Psychological Research Unit No.	Form	N	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$
Pilots in primary training	43G, 43I	3	CI616A	666	0.80	15.3	11.1	6.8	0.36	...
Pilots in basic training	43I, 43J	3	CI616A	401	.93	14.5	12.3	7.1	.15	...
Pilots through basic training	43I, 43J	3	CI616A	533	.70	14.5	10.9	7.0	.31	[2]0.34
Pilots in primary training	44C	3	CI616B	877	.90	36.3	29.1	10.5	.35	...
De	44E	3	CI616A	585	.92	16.2	13.3	6.5	.22	[3].28
De	44F	3	CI616B	479	.90	34.2	27.7	10.8	.30	[3].41
De	44H	2	CI616B	504	.92	34.54	29.12	10.15	.27	[3].34
De	44H	2	CI616B	406	.93	32.17	32.02	9.12	.12	.18
De	44I	1, 2, 3	CI616B	3,146	.84	32.71	29.11	10.31	.19	[4].29
De	44J	3	CI616B	1,676	.89	29.39	25.76	9.28	.20	[5].33
WASPS	44-W-8	...	CI616B	104	.61	28.00	20.44	10.29	.46	...

[1] Form CI616A consists of 30 items; Form CI616B consists of 60 items.
[2] Assumed unrestricted stanine standard deviation not reported.
[3] Assuming an unrestricted stanine standard deviation of 1.85.
[4] Assuming an unrestricted stanine standard deviation of 1.90.
[5] Assuming an unrestricted stanine standard deviation of 1.83.

TABLE 19.6.—*Validity data for Instrument Comprehension, CI616B, using various criteria*

Group	Unit	Criterion	N	r[1]
Navigator trainees	PRU No. 3	Grades in flight missions	200	0.16
De	PRU No. 3	Grades in ground missions	200	.28
De	PRU No. 3	Weighted average grades	200	.28
Armament trainees[2]	MPEU Nos. 1-10	Average grades	269	.05
Officer candidates	MPEU No. 5	Eighth week academic average	343	.25
De	MPEU No. 5	Eighth week ratings	343	.08
De	MPEU No. 5	Twelfth week ratings	343	.02

[1] Product-moment correlations.
[2] In Lowry Field classes 34-44A and 35-44A.

703320—47—32

TABLE 19.7.—*Validity of items of Instrument Comprehension based upon graduation-elimination of pilots in primary training (in class 44H; N=704; p_e=.85)*

Form	Mφ	SDφ	Range of φ	
			Low	High
CI615B	0.11	0.07	−0.02	0.24
CI616B	.08	.16	−.01	.22

the fact that it was not preceded by part I. This change was accomplished in September 1944, when part II was given a new set of directions and was continued in the battery as CI616C.

Its contribution was at least two-fold. First, it was a substantial member in the family of valid group tests. Second, in factor analysis studies, it helped to shed light on the question of why certain other tests were valid, thus illuminating the way for further test construction.

That Instrument Comprehension's validity was due in great part to its measurement of the spatial-relations factor was the subsequent revelation of factor analysis. To realize what a gain in insight this was, we have only to refer back to the original rationale for the test and note the six different abilities it was supposed to measure. As things turned out, it measures only one ability related to the six (visualization) and, in addition, it measures several other quite differently defined abilities. This gain in understanding pointed the way to a new line of research and test construction, aimed at developing printed tests of spatial ability. Results show that no other new or unique factor was needed to account for the validities of approximately 0.20 for part I and 0.30 for part II (see table 28.18).

Aerial Orientation, CP520A *

Instrument Comprehension sampled the ability to interpret position from instrument readings, but no measure had been constructed to test the ability to interpret position from outside cues visible from the pilot's position in the cockpit. It was hypothesized that the latter ability is more important to the pilot, both in training and in combat flying. Aerial Orientation was designed to measure reaction to external cues seen from within the cockpit. It was intended, due to the necessity of the pilot's perceiving plane attitude on the basis of spatial cues rather than on the basis of intellectual interpretations, that Aerial Orientation should contain more spatial and less verbal and reasoning factor content than Instrument Comprehension.

Description.—For each item, a cockpit view is shown diagrammatically on the left-hand side of the page, set opposite photographs of a model airplane in five different attitudes. The cockpit views are drawings representing patterns of land and water that might be seen from an airplane maneuvering over a point on a coastline. The examinee's problem is to

* Developed at Psychological Research Unit No. 3. Chief contributor: S/Sgt. Wayne S. Zimmerman.

match the cockpit view with the airplane from which that view would be seen.

(1) *Internal characteristics.*—The directions contain two recorded but unscored sample items. Part I contains 30 scored items, and part II, 28.

(2) *Administration.*—Six minutes are allowed for administration of the directions, 10 minutes for doing items in part I, and 8 minutes for doing items in part II, making a total testing time of 24 minutes.

A sample item is shown in figure 19.3. Following are part of the directions:

This is a test of your ability to visualize the relationship between a plane and the territory over which it flies.

Look at the sample problem. The picture on the left represents the landscape as it would appear to the pilot of one of the five planes on the right. Your task is to select the plane from which the pilot would see this view when he looks directly over the nose of his plane. Notice that in the group of five pictures the ocean is on the right and land is on the left, with the coast line directly under the plane.

Study the sample problem and select the plane from which the pilot would see the landscape as it appears in the left picture.

According to the picture on the left, the correct plane is flying level, unbanked, and headed directly over the land. Thus D is the correct answer because from this plane the pilot would see only land, and the horizon would appear to be level.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results.—The data given below all are based upon examinees tested at Psychological Research Unit No. 3.

(1) ***Distribution statistics.***—Typical distribution statistics are given in table 19.8. The distribution curves are symmetrical and considerably flatter than normal.

TABLE 19.8.—*Distribution constants for Aerial Orientation, CP520A based upon samples of classified pilots*

N	M	SD
[1]282	37.5	12.7
[2]610	41.7	12.3

[1] In classes 44G and 44H.
[2] In class 44H.

(2) *Internal consistency.*—The degree of homogeneity of the test is indicated by a mean internal-consistency phi of 0.42, a standard deviation of the phi distribution of 0.12, and a range of values from 0.14 to 0.60. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score of a group of 750 unclassified aviation students tested in July 1944.

(3) *Reliability coefficient.*—By correlating part I with part II, an estimated reliability coefficient of 0.84, corrected for length, was obtained. This figure is based on a sample of 443 classified pilots in class 44H.

(4) *Correlation between rights and wrongs.*—For a sample of 330 pilots tested in the period from July 4 to July 26, 1944, the correlation between rights and wrongs was −0.77. For another independent sample of 500 unclassified aviation students (tested early in 1945) the correlation was −0.75.

(5) *Difficulty.*—Based upon the responses of the above-mentioned sample of 750 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.63, corrected for chance, a standard deviation of 0.18, and a range from 0.14 to 0.98.

(6) *Test validity.*—Validation results based on several samples are given in table 19.9.

(7) *Item validity.*—Validation of items revealed a mean phi of 0.08, based upon the responses of 600 graduates and 104 eliminees from primary pilot training in class 44H. The standard deviation of phi values was 0.06, and the range was from −0.12 to 0.22.

Variations.—Form CI520B[7] is a lengthened version of the original test embodying two significant changes. First, the airplane model used in the photographs is more realistic in appearance, and its various attitudes are more easily perceived. Second, for the purpose of simplifying the problems and increasing the speed factor, only 45° and 90° turns, banks, and altitude changes (climbs and dives) are presented. The B form was fully developed but was not printed in booklet form, because it was developed very late in the program, when experimental testing ceased.

Form CI520C is made up of the items from part I of Form A and is divided into two equal parts.

Evaluation.—Aerial Orientation, CP520A, proved to be one of the most valid printed tests constructed for pilot as well as navigator selection. It would have added slightly to the total validity of the classification battery, even with Instrument Comprehension (the test it most closely resembles), as a member. This finding indicates that it either measures known valid factors better than other tests that are in the battery, or that it possesses a unique valid factor or factors, or both. Intercorrelations indicate that, as planned, Aerial Orientation, is probably a purer measure of a spatial factor than is Instrument Comprehension. The indications are that Aerial Orientation would do very well as a substitute for Instrument Comprehension in any alternate battery. Its extremely high navigator validity, however, indicates factorial complexity and somewhat dims its promise as a discriminating classification test.

Flight Orientation, CP528A [8]

The idea for Flight Orientation was proposed at the time Aerial Orientation was being developed. It was hypothesized (1) that the ability visually to maneuver an airplane as if from a position outside the cockpit is

[7] Developed at Psychological Research Unit No. 3. Chief contributor: Staff/Sgt. Wayne S. Zimmerman.

[8] Developed at Psychological Research Unit No. 3. Chief contributors: Pvt. Charles K. Ferguson and Staff/Sgt. Wayne S. Zimmerman.

TABLE 19.9.—*Validity data for Aerial Orientation, CP520A, graduation-elimination criterion*

Group	Class	Scoring formula	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$
Pilots in primary training	45C	Rights only	330	0.82	40.58	35.20	10.47	0.29	[1]0.44
Do	45C	Wrongs only	330	.82	13.74	15.87	8.41	−.14	[1]−.29
Do	44H; 44I	R−W/4	1,122	.87	41.23	35.72	12.24	.24	[1].29
Do	45C	R−W/4	330	.82	37.14	31.23	12.16	.28	[1].43
Navigators	([3])	Rights	478	.90	41.99	35.96	10.09	.31	[2].50
Do	([3])	Wrongs	478	.90	8.71	11.77	6.10	−.26	[2]−.41
Do	([3])	R−W/4	478	.90	39.81	33.02	11.05	.32	[2].50

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Assuming an unrestricted stanine standard deviation of 1.87.
[3] Class unidentified. Tested in May 1944.

a manipulatory-visualization ability and (2) that the ability to imagine maneuvers taking place as if the examinee were within the cockpit is a spatial-orientation ability.

The Aerial Orientation test utilized cockpit views of outside terrain to be matched with depicted plane attitudes; the visualization-of-maneuvers tests involved only views of airplanes seen from a position outside of the cockpit (see ch. 12 for a discussion of these tests). Flight Orientation was designed to fulfill the requirements of the indicated variation—a test that would utilize only cockpit views of outside terrain. From hypotheses given above, it follows that Aerial Orientation should measure a combination of manipulatory-visualization and spatial-orientation abilities, while Flight Orientation should be a purer measure of the ability to orient in space.

Description.—Each item consists of two landscape photographs taken from the cockpit of an airplane. The first photograph represents the view seen by the pilot when the plane is in one given position. The second shows the view seen after a certain maneuver has been completed. The examinee must decide which of a number of maneuvers has taken place. The maneuvers include right turns, left turns, right rolls, left rolls, and climbs and dives. Answers are marked on a special, overprinted answer sheet to be described later.

(1) *Internal characteristics.*—Part I contains 3 recorded, but unscored, sample items and 47 scored items. Part II contains 50 scored items. In Part I, only one maneuver is represented as having been completed between any two photographs. In Part II, two combined maneuvers, such as a turn and a dive, are represented. In this part, two descriptive terms are required for each item in order to describe the action.

(2) *Administration.*—Eight minutes are allowed for Part I and 14 minutes for Part II. The directions require approximately 5 minutes, making a total testing time of 27 minutes.

A sample item for Part I is shown in figure 19.4. Following are part of the directions.

> This is a test of your ability to recognize change in flight position.
>
> Look at the sample problem. Two pictures are shown. The picture at the left shows a cockpit view. The picture at the right shows the cockpit view as it appears after a single maneuver. Your task is to determine the maneuver. The maneuver will be one of the following: Left or right turn; left or right roll; climb, up or down.
>
> The correct answer to sample problem 1 is "right roll."
>
> The answer to item 1 is marked correctly on the illustration of the answer sheet (shown in fig. 19.5).

(3) *Scoring.*—The scoring formula used is R−W/3.

Statistical results.—The analysis of this test is only partially complete. The data all are based upon examinees tested at Psychological Research Unit No. 3.

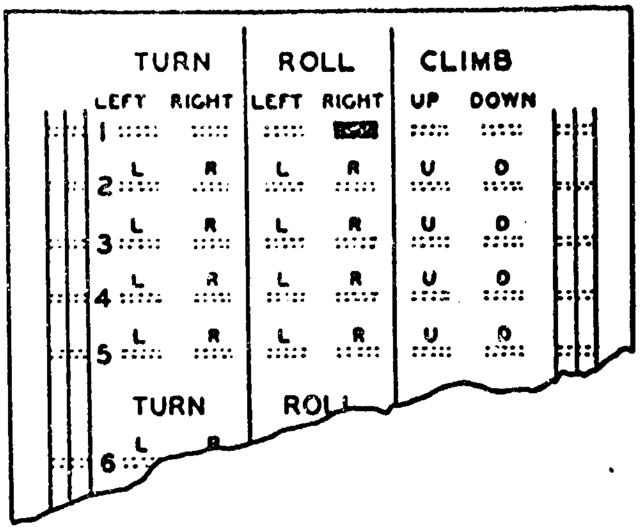


FIGURE 19.5
SECTION OF ANSWER SHEET OF FLIGHT
ORIENTATION, CP528A

(1) *Internal consistency.*—Analysis of the responses of one sample group yielded the internal-consistency data given in table 19.10.

TABLE 19.10.—*Internal-consistency data for Flight Orientation, CP528A, based upon a sample of unclassified aviation students*[1]

Part	Mφ	SDφ	Range of φ	
			Low	High
I	0.30	0.14	0.04	0.63
II	.37	.17	−.10	.74

[1] N=750 tested from September 26 to Nov. 22, 1944. The criterion is total rights score, parts I and II combined.

(2) *Reliability coefficient.*—Based on a sample of 502 unclassified aviation students tested from August 10 to October 6, 1944, a correlation of 0.78 (corrected for length) was secured between part I and part II of this test. Since the two parts are not completely comparable, this figure can be considered only as a very rough estimate of the reliability.

(3) *Difficulty.*—Based upon the responses of the above-mentioned sample of 750 unclassified aviation students, Parts I and II of the test yielded the mean proportions of correct responses given in table 19.11.

TABLE 19.11.—*Difficulty of items for Flight Orientation, CP528A*

Part	M_{cp}	SD_{cp}	Range of difficulty	
			Low	High
I	0.61	0.17	0.16	0.88
II	.54	.19	.10	.91

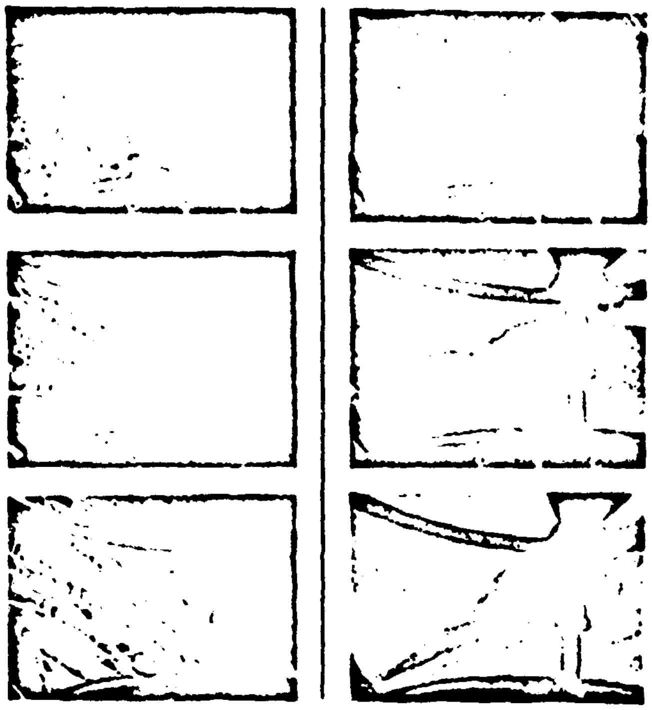


FIGURE 19.6
SAMPLE ITEMS O⁼ ST'CK AND RUDDER ORIENTATION. CP531A

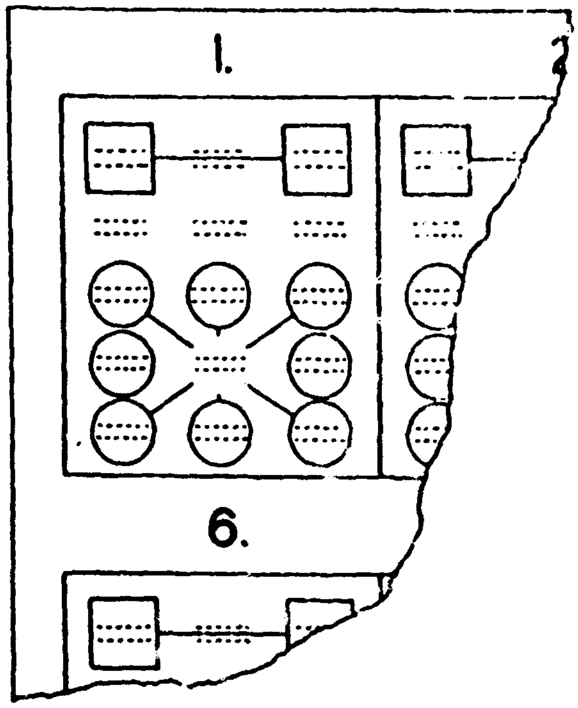


FIGURE 19.7
SECTION OF ANSWER SHEET OF STICK AND RUDDER ORIENTATION, CP521A

to a maneuver or a completed maneuver. For example, problem 1 shows an airplane banking to the right. Your task is to record on the special answer sheet the positions to which the stick and rudder would be moved to perform each maneuver.

For the problems in this test the stick and rudder are moved as follows:

For banks only—To bank left, move stick to left. To bank right, move stick to right.

For banked turns—To turn left, move both stick and rudder to left. To turn right, move both stick and rudder to the right.

For climbs and dives—To dive, push stick forward. To climb, pull stick back.

On your answer sheet indicate stick and rudder positions according to the diagram of the answer sheet and accompanying key. (See fig. 19.6.)

Reading downward, the series of three cockpit views would be seen from an airplane banking to the right. To maneuver the airplane in the direction the stick would be moved to the right. Therefore, right stick is correctly marked on the sample answer sheet.

(3) *Scoring.*—The scoring formula used is R−W/5.

Statistical results.—None are available.

Discrimination Reaction Time (Paper), CP634A [11]

The economy provided by printed tests administered to large groups is so great, compared with the more difficult and troublesome administrative requirements of apparatus tests, that there was spasmodic interest in the possibility of duplicating valid apparatus tests in the simpler printed form.

[11] Developed at Psychological Research Unit No. 1. Chief contributor: Staff/Sgt. Wayne S. Zimmerman.

Early in the program, little success was realized in duplicating the functions of apparatus tests in the form of printed tests. This was partially due to the fact that the factorial content of apparatus tests used in the classification batteries was not well known. It had been assumed that the validity of Complex Coordination, CM701A, Two-Hand Coordination, CM101A, Discrimination Reaction Time, CP611D, and other psychomotor tests included in the classification battery was due primarily to the motor-coordination elements measured. Consequently, attempts were made to maintain the conspicuous motor aspects of tests when printed material was employed. This naturally presented obstacles that were not easy to overcome. But after it was found that several apparatus tests were highly correlated with some printed tests, it was discovered that they actually measured spatial-relations, perceptual-speed, and visualization abilities as well as motor abilities. The result was encouraging, and the way was clear to attempt the desired duplications by means of printed tests.

Description.—Discrimination Reaction Time, CP634A, presents, on paper, patterns similar to those shown with red and green lights in the apparatus Discrimination Reaction Time test.[12] Black and white circles take the place of the colored lights. The examinee responds to the stimuli by marking in one of the four directions—up, down, left, or right—on a specially designed answer sheet. Each of the four directions for marking corresponds to the direction that the examinee would move in order to snap one of the four switches on the discrimination reaction time apparatus.

Parts I and II present patterns almost identical with the light patterns in the apparatus test. Parts III and IV call for a response to slightly more complex patterns. The stimulus is the arrangement of three circles, one white, one black, and one with a cross. The response is made in one of the same four directions—up, down, left, or right.

(1) *Internal characteristics.*—The directions for parts I and II include four unrecorded sample items and five recorded but unscored sample items. Part I contains 45 scored items, and part II contains 50 scored items. The directions for parts III and IV contain five recorded, but unscored, sample items. Part III contains 45 scored items, and part IV contains 50 scored items. All of the items in each part are included on a single page. Each item is located on the page in the position that corresponds to the position of its item number on the answer sheet. The items in parts III and IV, calling for a response to a three-circle pattern instead of a two-circle pattern, are more difficult than those in parts I and II.

(2) *Administration.*—Answers are marked directly on the special, overprinted, answer sheet described later. One minute and 15 seconds are allowed for part I, 1 minute for part II, 1 minute and 30 seconds

[12] For a brief description of this test see page 804. For a detailed description see Report No. 4.

for part III, and 1 minute and 15 seconds for part IV. Administration of directions and sample items takes 5 minutes, making a total testing time of approximately 10 minutes.

Examinees are instructed to make only a small check, short line, or other quick mark to indicate their answers. They are told that time will be allowed later to fill in the spaces completely. At the end of parts I and II they are instructed to close their test booklets and to go back and fill in the answer spaces for both completed parts. At the end of the test the same instructions are again given for parts III and IV. This feature was later discarded.

Correctly marked sample items for parts I and II are shown in figure 19.8. Following is part of the directions for parts I and II:

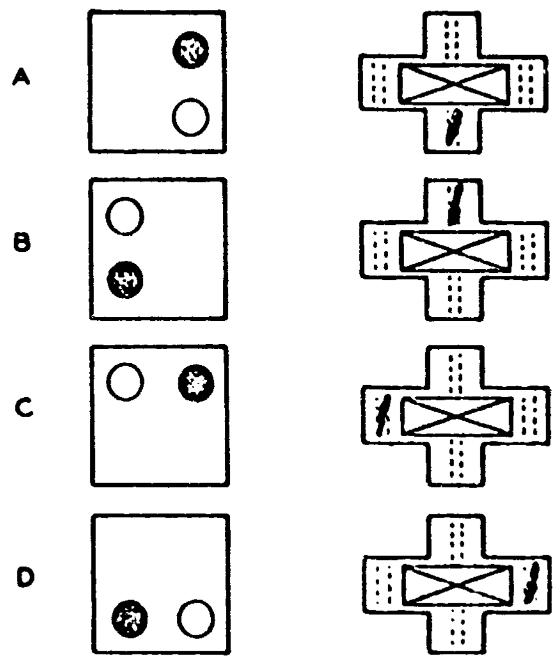


FIGURE 19.8
SAMPLE PROBLEMS OF DISCRIMINATION REACTION TIME (PAPER) - CP634A - PARTS I & II AND CORRECTLY MARKED ANSWER SHEETS

This is a test of speed of reaction to a signal. The signal will be an arrangement of a black and a white circle. There are only four arrangements of the circles, and four ways to mark your answer sheet. Look at the sample problem below and the corresponding illustrations of the correct ways to mark your answer sheet.

A. When the white circle is below the black circle, mark the lower space in the cross.

B. When the white circle is above the black circle, mark the upper space in the cross.

C. When the white circle is to the left of the black circle, mark the space to the left.

D. When the white circle is to the right of the black circle, mark the space to the right.

Following is a part of the directions for parts III and IV:

In this part of the test the signal will be the arrangement of three circles instead of two. (See fig. 19.9.)

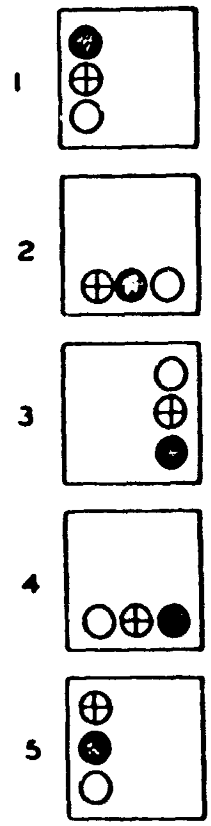


FIGURE 19.9
SAMPLE PROBLEMS OF DISCRIMINATION REACTION TIME (PAPER), CP634A
PARTS III & IV

1. If the black circle is on the outside, mark in its direction.
2. If the black circle is in the center, mark in the direction of the white circle.

Remember:

Black outside, mark toward black.
Black center, mark toward white.

Work sample items 1 through 5.

The correct answers should be marked as follows:

Item 1, upper space.
Item 2, right space.
Item 3, lower space.
Item 4, right space.
Item 5, lower space.

(3) *Scoring.*—The scoring formula used is R−W/3.

Statistical results.—None are available.

Directional Marking, CP533A [13]

This test was designed to make a decisive examination of the hypothesis that the space I factor is a directional-discrimination ability. There was an attempt, in construction, to include movements in three dimensions. On a flat sheet of paper, such as an answer sheet, however, only two-dimensional movement is possible. Up and down and left and right were selected as the descriptive terms in these two dimensions. Since the third dimension of depth could not be used, the closest substitute suggested was to use the terms near and far, which are ordinarily descriptive of depth, but which were made to apply to the two-dimensional surface as limitations on the movements in the flat-surface dimensions.

Description.—Each item consists of four, verbally stated directions of movement, and each of these four statements describes the position of an answer space that is within a square, printed on the answer sheet, containing 25 answer spaces. The center space in each square or box is covered by a solid black circular spot, which is designated as the starting point for each of the responses to items. The examinee's task is to place four marks in each box at the distances and directions from the center circle that are described in the four verbally stated problems. Figure 19.10 shows one correctly marked sample item of the test.

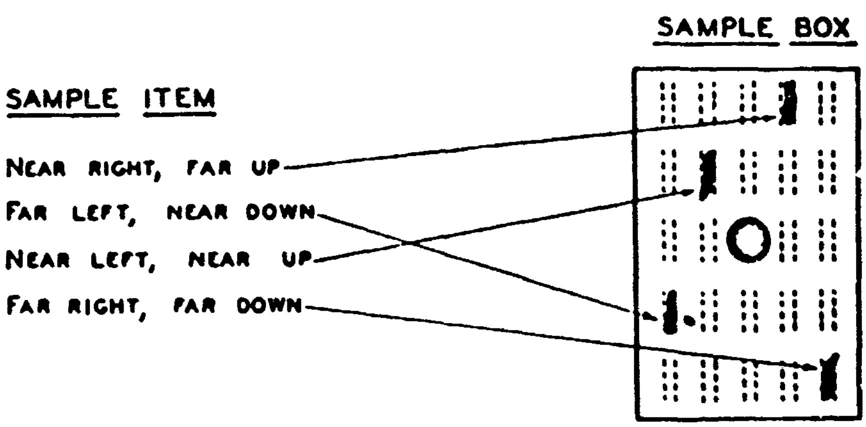


FIGURE 19.10
SAMPLE ITEMS WITH CORRECTLY MARKED ANSWERS FOR DIRECTIONAL MARKING, CP533A

(1) *Internal characteristics.*—All of the answers are marked on one side of an answer sheet. Since there are 25 answer spaces inclosed within each item box and 750 answer spaces on one side of an answer sheet, there are 30 item boxes. One of these is used for the sample problem, 14 are used for part I and the remaining 15 are used for part II. Thus, the directions contain 4 recorded but unscored answers, part I contains 56 recorded and scored answers, and part II contains 60 scored answers.

[13] Developed at Psychological Research Unit No. 3. Chief contributors: Tech./Sgt. Paul C. Davis, Pfc. Harold H. Kelley, Capt. Lloyd G. Humphreys, and Staff/Sgt. Wayne S. Zimmerman.

(2) *Administration.*—Three minutes are allowed for recording answers in part I, and 3 minutes and 15 seconds are allowed in part II. Administration of directions and explanation and recording of the sample problem takes 5 minutes, making a total testing time of 11 minutes.

Following is part of the directions:

Each item in this test consists of four problems; each problem requires you to place a mark in a box * * * The box appears on your separate answer sheet and has a dot in the center surrounded by spaces for the marks.

Your task is to place the four marks at the distances and directions from the dot which are indicated by the following instructions:

1. Near right means one space from the dot to the right; near left, one space to the left.
2. Far right means two spaces from the dot to the right; far left, two spaces to the left.
3. Near up means one space up from the dot; near down, one space down.
4. Far up means two spaces up from the dot; far down, two spaces down.

(3) *Scoring.*—The scoring formula used is R − W/4.

Statistical results.—None are available.

Evaluation of Directional Discrimination Tests

The question of whether there is a single factor in common to all tests in this group is still unanswered, although evidence suggesting such a conclusion is considerable. The study of tests in this area has been particularly challenging, (1) because the spatial factor is highly valid for both pilot and navigator selection, and (2) because it has been difficult to develop a pure measure of the factor. No test constructed as yet has demonstrated a loading substantially greater than 0.50 for the spatial-relations factor. The validity of this factor for all three air-crew assignments justifies maximal effort to enlarge this loading and to purify its measuring instrument.

If the hypothesis that a directional-discriminational ability is the important aspect of the spatial-relations factor, then we should expect Directional Marking or Discrimination Reaction Time (paper) to show the greatest purity. Directional Marking is crucial for the hypothesis because it removes the element of visually perceived spatial arrangements from consideration by presenting the stimuli verbally.

POSITIONAL DISCRIMINATION TESTS

The latest factor study referred to in the introduction to this chapter proved to be important not only because it demonstrated the spatial nature of the factor that had been in question, but also because it pointed to the existence of a second spatial factor. That this second factor had not emerged until this particular analysis is attributable to the fact that the Hands test had not been incorporated in any of the previously studied batteries. Until its inclusion, presumably, other tests that appreciably measured the ability had been too few or too weak in the ability to make its presence known. Other tests having this factor in common are Flags, Figures and Cards, CP512A, and Cubes, CP512A.

Several tentative names were proposed for this variable. These included: "hands" space, rotational space, spatial empathy, and positional discrimination. The term "spatial empathy" was suggested because informal introspective reports of some psychologists indicated that in solving the items one may "project" himself into the test objects or into a more favorable posture from which to judge the positions of the objects.

The following tests are designated "Positional Discrimination" tests and are treated together as the second subarea of tests of a spatial chapter. The three tests described are revisions of those mentioned above and are named Positional Orientation, CP526A, Object Identification, CP521A, and Object Recognition, CP523A. They are discussed in the order of their development.

Object Identification, CP521A [14]

This is a revision of Thurstone's Flags test. It was adapted for the purpose of measuring and studying the hypothesized ability to manipulate images in space. This revision was further motivated by the promising validity of Thurstone's Flags test which had been adapted by the AAF in a systematic study of perceptual tests. Face validity was incorporated into one section of the revised form of the test by using, instead of flags, silhouettes of military vehicles.

Description.—Silhouettes of planes, trucks, guns, tanks, and ships are presented in part I, and flags are used in part II. The examinee's task is to select, from five illustrations, those that show the same side of an object as that shown in a key illustration. Some are turned over and are, therefore, incorrect answers.

(1) *Internal characteristics.*—The directions contain two recorded but unscored sample items. Part I contains 28, and part II 30 scored items presented in line drawings.

(2) *Administration.*—Directions consume 3 minutes, and 7 and 6 minutes are allowed to complete parts I and II respectively, making a total testing time of 16 minutes.

One sample item from part I of the test is shown in figure 19.11. Following are part of the directions:

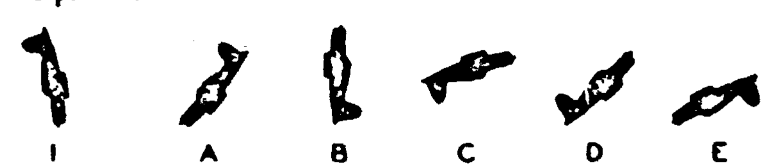


FIGURE 19.11
SAMPLE ITEM OF OBJECT IDENTIFICATION, CP521A

This is a test of your ability to identify objects in silhouette.

You will see rows of six silhouettes. Your task will be to compare the five silhouettes labeled A, B, C, D, and E with the first silhouette in each row.

[14] Developed at Psychological Research Unit No. 3. Chief contributor: Lois G. Wright.

Some of the five silhouettes are the same as the first one in the row, but have been slid around into different positions.

Others of the five silhouettes are different from the first one in the row; that is, they have been turned over and could not be made to fit the first silhouette by simply sliding them around on the page.

Your problem will be to decide which of the five silhouettes in each row match the silhouette on the left.

Plane silhouettes A, B, and C are the same as silhouette number 1. Therefore A, B, and C are the correct answers for item number 1.

(3) *Scoring.*—The scoring formula used is R−W+100.

Statistical results.—Unless specifically noted to the contrary, the data below are for examinees tested at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 19.12. The distribution curves are considerably negatively skewed and somewhat flatter than normal.

TABLE 19.12.—*Distribution constants for Object Identification, CP521A based upon samples of classified pilots*

N	M	SD
[1]282	184.7	21.4
[2]1,222	186.7	24.0

[1] In classes 44G and 44H.
[2] In class 44H.

(2) *Reliability coefficient.*—Three samples yielded the estimates of reliability given in table 19.13. Table 3.1 presents additional reliability data.

TABLE 19.13.—*Alternate-forms (part I v. part II) reliability coefficients for Object Identification, CP521A*

Group	N	r'_{II}	r_{II}
Pilots[1]	348	[2]0.66	0.80
Unclassified aviation students[3]	500	[2].65	.79
Unclassified aviation students[3]	500	[4].60	.75
Unclassified aviation students plus airplane mechanics[3]	448	[2].64	.78

[1] In class 44H.
[2] Part II administered immediately after part I.
[3] Tested at Medical and Psychological Examining Units Nos. 6 and 8 early in 1945.
[4] Part II administered approximately 4 hours after part I.

TABLE 19.14.—*Correlation between rights and wrongs for Object Identification, CP521A*

Group	N	Part	r_{RW}
Unclassified aviation students[1]	500	I	−0.28
Do[1]	500	II	−.22
Pilots[2]	189	I and II	−.40
Navigators[3]	1,257	I	−.28
Do[3]	1,257	II	−.28
Do[3]	1,257	I and II	−.30

[1] Tested early in 1945. Exact testing dates not identified.
[2] Tested in July, 1944.
[3] Tested at Psychological Research Unit No. 1 in June 1944; at Psychological Research Unit No. 2 in April, 1944; and at Psychological Research Unit No. 3 in March 1944.

(3) *Correlation between rights and wrongs.*—The data are shown in table 19.14.

(4) *Factorial composition.*—The most significant loadings for Thurstone's tests, Flags, Figures, and Cards, CP512A, the earlier version of Object Identification, CP521A, are in the spatial-relations (0.43), space II (0.42), and perceptual-speed (0.31) factors. The communality is 0.54 which almost equals the test reliability. For a fuller picture of the factorial composition of this test see Appendix B.

(5) *Test validity.*—Validation results based on several samples are given in table 19.15.

(6) *Item validity.*—Validation of items revealed a mean phi of 0.08, based upon the responses of 600 graduates and 93 eliminees from primary pilot training in class 44I. The standard deviation of phi values was 0.08 and the range was from −0.12 to 0.23.

Evaluation.—Flags, Figures, and Cards, the test from which Object identification was derived, was factor analyzed, but the derived form was not. Forty-six percent of the total variance of Flags is attributable to three factors. Eighteen percent of the total variance is found in the spatial-relations factor, 18 percent in the space II, and 10 percent in the perceptual-speed factor. The known factors in the Flags test exactly account for its average pilot validity of 0.24, which allows a small loading of 0.05 for the factor space II in the pilot criterion (see table 28.17).

Part II of Object Identification showed somewhat higher pilot validity than part I in one sample but not in another. It would be of interest to treat parts I and II as two separate variables when factor analyzing them.

The substantial navigator validity calls for investigation to determine its source.

Object Recognition, CP523A [15]

This is a revision of Thurstone's Cubes test. It was adapted for the purpose of obtaining another measure of the hypothesized ability to manipulate images in space. Solving problems in this test seemed to call for visualization in three dimensions from presentation in only two dimensions. In Thurstone's factor analyses, Cubes appeared heavily saturated with his space factor (factor S), a factor that, if we are to accept Thurstone's description, seems more closely to resemble visualization. A reanalysis of Thurstone's data showed a rather clear separation of space and visualization.

For the purpose of adding face validity to the items, insignia of the United States Army were substituted for the nonsense symbols of Thurstone's cubes.

Description.—An item consists of an illustration of a single cube (key cube) in a row with five other (alternate) cubes. On each of the six

[15] Developed at Psychological Research Unit No. 3. Chief contributors: Cpl. Albert A. Canfield Jr. and Staff/Sgt. Benjamin Fruchter.

703320—47—33

TABLE 19.15.—*Validity data for Object Identification, CP521A, graduation-elimination criterion*

Group	Class	Part	Scoring formula	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$
Pilots in primary training	44H	I	R−W	783	0.88	47.90	44.30	14.35	0.13	[1]0.16
Do	44H	II	R−W	783	.88	41.45	35.85	12.70	.23	[1].26
Do	44H–44I	I–II	R−W±100	[2]1,122	.87	187.93	180.08	24.77	.17	[3].22
Do	[4]	I	Rights	189	.82	53.76	50.97	10.68	.15	[5].22
Do	[4]	I	Wrongs	189	.82	2.50	3.38	3.48	−.14	[5]−.16
Do	[4]	II	Rights	189	.82	46.44	44.00	9.71	.14	[5].20
Do	[4]	II	Wrongs	189	.82	2.54	3.15	2.87	−.12	[5]−.18
Do	[4]	I–II	R−W	189	.82	95.70	88.47	20.58	.20	[5].27
Navigators	[6]	I	Rights	1,257	.92	54.90	51.92	10.61	.15	[5].30
Do	[6]	I	Wrongs	1,257	.92	2.11	2.67	3.42	−.09	[5]−.16
Do	[6]	II	Rights	1,257	.92	48.60	45.58	10.55	.16	[5].30
Do	[6]	II	Wrongs	1,257	.92	2.41	3.18	3.01	−.14	[5]−.25
Do	[6]	I–II	Rights	1,257	.92	103.50	97.50	18.74	.17	[5].33
Do	[6]	I–II	Wrongs	1,257	.92	4.52	5.85	5.37	−.14	[5]−.25
Do	[6]	I–II	R−W	1,257	.92	98.98	91.65	21.01	.19	[5].35

[1] Assuming an unrestricted stanine standard deviation of 1.79.
[2] Includes the preceding sample of 783 pilots tested with parts I and II.
[3] Assuming an unrestricted stanine standard deviation of 1.83.
[4] Class unidentified. Sample tested in July 1944.
[5] Assuming an unrestricted stanine standard deviation of 2.00.
[6] Class unidentified. Sample consists of examinees tested at Psychological Research Unit No. 1 in June 1944, at Psychological Research Unit No. 2 in April 1944 and at Psychological Research Unit No. 3 in March 1944.

sides of a cube are different military insignia, but only three sides show in the illustrations. The examinee's task is to determine whether or not each of the five alternate cubes could represent the key cube; that is, having every side with the same insignia in proper relationships. The alternate cubes are either rotated or turned from the original position, have different insignia, or have the insignia placed in different relationships than on the key cube.

(1) *Internal characteristics.*—The directions contain one sample row of cubes, and parts I and II each contain 10 key cubes. The sample requires 5 judgments and each part requires 50 judgments, 5 for each key cube.

(2) *Administration.*—Two minutes are allowed for administration of the directions and sample items, 13 minutes for part I, and 12 minutes for part II, making a total testing period of 27 minutes.

A sample is shown in figure 19.12. Following are part of the directions:

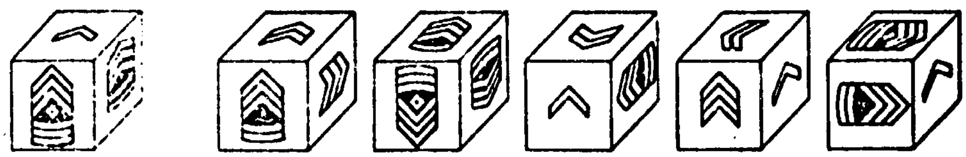


FIGURE 19.12
SAMPLE ITEM OF OBJECT RECOGNITION,
CP523A

This is a test of your ability to visualize change of position. You will be shown drawings of cubes. Each cube has six sides, and each side has a different military insignia. Look at the sample problem.

The cube at the far left is the key cube. Your task is to select from the five cubes at the right the ones that could represent the key cube turned to a different position.

Cubes C and D are correct answers. Both could be the key cube turned to a different position.

(3) *Scoring.*—The a priori scoring formula is R−W. Most of the statistics reported on the test, however, are for rights and wrongs scored separately.

Statistical results.—The available data all are based upon examinees tested at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 19.16. The distribution curves for rights scores are moderately negatively skewed and considerably flatter than normal, and for wrong scores are positively skewed and somewhat flatter than normal.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi (based upon the total group) of 0.52, a standard deviation of the phi distribution of 0.14, and a range of values from 0.05 to 0.84. These statistics are based upon the responses of the highest 27 percent and the lowest 27 percent in total

TABLE 19.16.—*Distribution constants for Object Recognition, CP523A*

Group	Part	Scoring	N	M	SD
Classified pilots	I and II	Rights	[1]1,330	29.9	7.[illegible]
Do	...do	Wrongs	[1]1,330	11.5	5.4
Unclassified aviation students	I	Rights	[2]500	16.4	4.3
Do	II	Rights	[2]500	13.9	4.3
Do	I and II	Rights	[2]500	29.2	7.9
Do	...do	Wrongs	[2]500	11.3	5.5

[1] In class 44I.
[2] Testing dates unidentified.

score of a group of 750 unclassified aviation students tested in May 1944. Contrary to usual practice, it will be noted, statistics are based on "total group." The reason for this is that total number of attempts for any one cube cannot be determined, since the examinee responds to each cube by marking or *not* marking each appropriate answer-space.

(3) *Reliability coefficient.*—One sample yielded the estimates of reliability given in table 19.17.

(4) *Correlation between rights and wrongs.*—For a sample of 849 pilots tested in the period from May 9 to July 10, 1944, the correlation between correct and incorrect responses was −0.22.

TABLE 19.17.—*Alternate-forms (part I vs. part II) reliability coefficients for Object Recognition, CP523A, based upon a sample of 500 unclassified aviation students*[1]

Score	r'_{tt}	r_{tt}
Rights	0.72	0.84
Wrongs	.62	.76

[1] Testing dates unidentified.

(5) *Difficulty.*—Based upon the responses of the above-mentioned sample of 750 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.69, corrected for chance, with a range from 0.50 to 0.86 and a standard deviation of 0.10.

(6) *Factorial composition.*—The most significant loadings for Thurstone's Cubes, CP512A, the earlier form of object recognition, were found in the spatial-relations (0.41), perceptual-speed (0.31), general-reasoning (0.26), space II (0.25), and visualization (0.20) factors. The communality was 0.53 to be compared with a reliability of 0.68. For a fuller picture of the factorial composition of this test see Appendix B.

(7) *Test validity.*—Validation results are given in table 19.18.

TABLE 19.18.—*Validity data for Object Recognition, CP523A, based upon graduation-elimination of pilots in primary training*

Score	N_t	p_s	M_s	M_e	SD_t	r_{bis}	${}_c r_{bis}$[1]
Rights	[2]1,330	0.86	29.28	27.16	7.53	0.15	0.23
Wrongs	[2]1,330	.86	11.36	12.71	5.45	−.13	−.18
Rights	[3]849	.78	31.67	29.41	7.89	.17	.26
Wrongs	[3]849	.78	10.83	10.87	5.15	−.01	−.07
R − W	[3]849	.78	20.84	18.54	10.33	.13	.24

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44I; tested in March 1944.
[3] Class unidentified. Sample tested from May 9 to July 10, 1944.

(8) *Item validity.*—Validation of items revealed a mean phi of 0.02, based upon the responses of 687 graduates and 112 eliminees from primary training in class 44I. The standard deviation of phi values was 0.05 and the range was from −0.08 to 0.31.

Evaluation.—Cubes, the predecessor of Object Recognition, was subjected to factor analysis, but the adapted form was not. The factorial picture is comparatively complex. The greatest portion of the total variance of the test on a single factor is 17 percent, which is attributable to the spatial-relations factor. Other factors showing some influence are perceptual speed with 10 percent, general reasoning with 7 percent, space II with 6 percent, and visualization with 4 percent of the total variance. These factors almost exactly account for the pilot validity of this test.

Both Cubes and Object Recognition showed moderate validity for pilots, but they failed to measure any function of known validity as well as other tests already in the classification battery. It may be expected that Object Recognition, when factor analyzed, will show even more perceptual content than Cubes due to the fact that identifying and comparing military insignia, such as those illustrated, is perceptually more difficult than the simpler and more obvious type of symbol used in Thurstone's Cubes test.

Position Orientation, CP526A [14]

This is an adaption of Thurstone's Hands test. The validity of Hands on a sample of 927 pilots was 0.26. On the same sample its correlation with the pilot stanine was only 0.17, which indicated that the test has substantial unique variance to offer. A revision was planned which would be more reliable and more adaptable to the group-testing procedures used in the AAF. The revision was also to include other right-left members of the body as well as hands. Validation on new samples was proposed in order to verify the validity of the Thurstone test.

Description.—Each item shows five drawings of right or left hands, arms, legs, eyes, or feet. The examinee's task is to determine quickly whether a drawing represents a right or a left member of the body. Parts I and II of the test show only hands while parts III and IV show hands, arms, legs, and feet.

(1) *Internal characteristics.*—The directions for part I contain four sample items. Each item has five illustrations, each calling for a response, making a total of 20 possible answers. Part I contains 26 items, making a total of 130 recorded and scored answers; part II calls for 150 answers, part III for 180, and part IV for 150. Directions and sample items for part I require five minutes, and the directions and four sample items for parts III and IV require an additional 3 minutes. The testing

[14] Developed at Psychological Research Unit No. 3. Chief contributors: Cpl. James B. Ferguson and Lt. Leon I. Hellman.

time for parts I and II is 7 minutes each, and for parts III and IV 7½ minutes each, making a total testing time of approximately 37 minutes.

A sample item from part I is shown in figure 19.13. Following are parts of the directions:

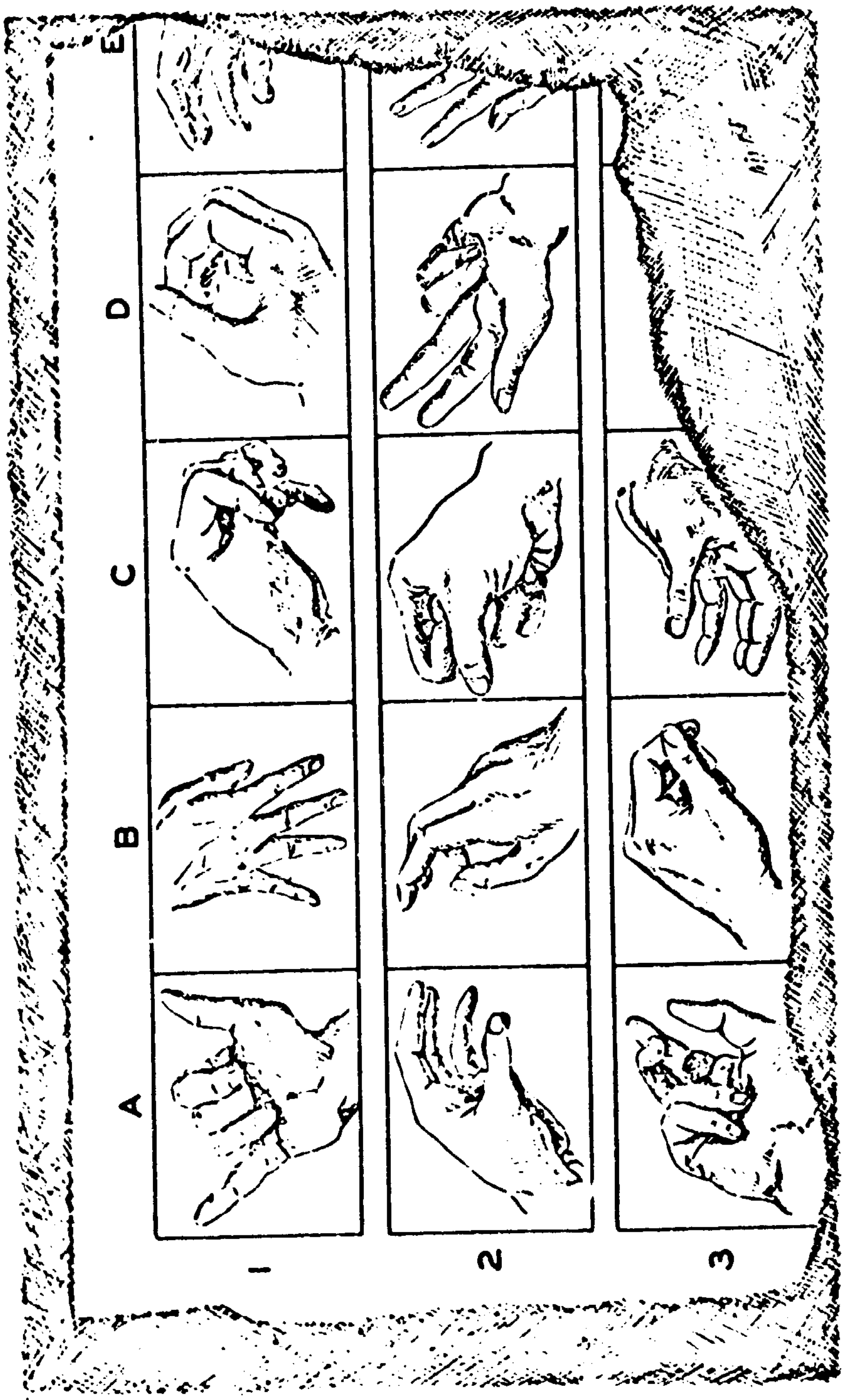


FIGURE 19.13
SAMPLE PROBLEMS OF POSITION ORIENTATION,
CP526A

In this test you will be shown a series of left and right hands in various positions. Your problem is to determine which of these are left hands and which are right hands. If the hand shown is a left hand, mark your answer in the answer sheet column labeled left. If it is a right hand, mark your answer in the answer sheet column labeled right. (See the illustration of the special answer sheet in fig. 19.14.)

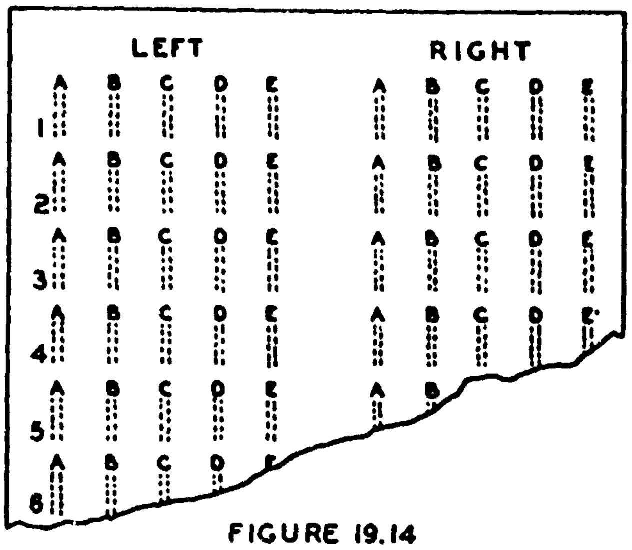


FIGURE 19.14
SECTION OF ANSWER SHEET OF POSITION ORIENTATION, CP526A

For example, look at problem 1. The hand shown in picture A is a right hand. Therefore, mark the space under A in the column headed right opposite problem 1 on your answer sheet. The hand in picture B is also a right hand, so the space under B must be marked in the column headed right.

(3) *Scoring.*—The scoring formula is R−W.

Statistical results.—The data are fairly complete on this test and are based upon examinees tested at Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Typical examples of distributions are given in table 19.19. The distribution curves are approximately symmetrical and considerably flatter than normal.

TABLE 19.19.—*Distribution constants for Position Orientation, CP526A, based upon a sample of 578 unclassified aviation students*[1]

Parts	M	SD
I and II	187.0	45.5
III and IV	164.6	46.8

[1] Tested in October 1944.

(2) *Reliability coefficient.*—By correlating part I and part II, an estimated reliability coefficient of 0.83, corrected for length, was obtained. This figure is based on a sample of 500 unclassified aviation students tested from August 10 to October 6, 1944.

(3) *Factorial composition.*—The only substantial loading (0.46) for the test Hands, CP512A, the earlier version of Position Orientation, is

on the space II factor. The communality is 0.35. For a fuller picture of the factorial composition of this test see Appendix B.

(4) *Test validity.*—Validation results based on several samples are given in table 19.20.

TABLE 19.20.—*Validity data for Position Orientation, CP526A, based upon graduation-elimination of pilots in primary training*

Part	Scoring formula	N_t [1]	p_s	M_s	M_e	SD_t	r_{bis}	$_cr_{bis}$ [2]
I	Rights	847	0.80	93.28	88.41	22.04	0.13	0.18
II	....do....	847	.80	98.63	95.19	25.91	.08	.15
III	....do....	847	.80	87.04	84.01	22.13	.08	.11
IV	....do....	847	.80	91.95	89.27	24.43	.04	.10
I	Wrongs	847	.80	4.02	4.61	3.69	−.09	−.12
II	....do....	845	.80	5.29	6.58	4.86	−.15	−.20
III	....do....	847	.80	7.06	8.67	5.02	−.18	−.22
IV	....do....	847	.80	10.06	11.12	6.60	−.09	−.13
I and II	Rights	847	.80	191.91	183.60	44.29	.11	.18
I and II	Wrongs	845	.80	9.31	11.19	7.47	−.14	−.19
I	Rights	294	.83	101.44	92.00	21.60	.16	.21
II	....do....	294	.83	102.80	90.80	23.57	.26	.28
III	....do....	294	.83	93.50	84.14	21.22	.25	.30
IV	....do....	294	.83	93.46	82.94	22.91	.26	.30
I	Wrongs	294	.83	6.11	7.30	5.29	−.13	−.16
II	....do....	294	.83	7.37	7.44	6.29	−.01	−.01
III	....do....	294	.83	9.28	9.16	6.05	.01	−.04
IV	....do....	294	.83	11.33	11.78	6.99	.04	−.08
I and II	Rights	294	.83	203.24	182.80	42.59	.23	.26
III and IV ..	....do....	294	.83	186.96	167.08	41.91	.26	.31
I and II	Wrongs	294	.83	13.48	14.74	9.97	−.07	−.10
III and IV ..	....do....	294	.83	20.61	20.94	11.68	−.02	.04
I	R−W	294	.83	95.33	84.70	22.79	.18	.24
II	R−W	294	.83	94.43	83.36	25.08	.25	.27
III	R−W	294	.83	84.22	4.98	22.77	.23	.29
IV	R−W	294	.83	82.13	71.16	24.29	.25	.39
I and II	R−W	294	.83	189.76	168.06	44.70	.23	.27
III and IV ..	R−W	294	.83	166.35	146.14	44.05	.26	.29

[1] Samples of 847 or 845 were tested in the period June 12 to July 12, 1944; the sample of 294 was tested in October 1944.

[2] Assuming an unrestricted stanine deviation of 2.00.

(5) *Intercorrelations.*—Part-score intercorrelations are given in table 19.21. The correlation between rights (parts I and II) and wrongs (parts I and II) was −0.13. For another sample of 294 pilots tested in October 1944, the correlation was −0.10. For this same sample, for the score on parts III and IV, the correlation between rights and wrongs was −0.05.

TABLE 19.21.—*Part-score intercorrelations for Position Orientation, CP526A, (N=847 classified pilots[1])*

Part	Score	1	2	3	4	5	6	7	8
1. I	Rights	...	0.70	0.64	0.64	−0.10	−0.11	−0.04	0.04
2. II	....do....	0.70	...	.68	.71	−.08	−.11	−.06	.04
3. III	....do....	.64	.68	...	.83	−.09	−.04	−.11	.02
4. IV	....do....	.64	.71	.83	...	−.06	−.03	−.06	−.05
5. I	Wrongs	−.10	−.08	−.09	−.06	...	.52	.45	.38
6. II	....do....	−.11	−.11	−.04	−.03	.52	...	.46	.37
7. III	....do....	−.04	−.06	−.11	−.06	.45	.46	...	.64
8. IV	....do....	.04	.04	.02	−.05	.38	.37	.64	...

[1] Tested in the period June 12 to July 12, 1944.

Evaluation

The validity coefficient of Position Orientation, parts I and II, was lower (0.18 on the largest sample reported) than the figure obtained on the original validation of the Hands test (0.26). Parts III and IV were even less valid. Whether this difference in validities is due to factorial differences between the original Thurstone form and the revised form of tests cannot be determined from present data. The biserial r's for the smaller sample were clearly in line with the 0.26 for Thurstone's version, however, so the discrepancy may be a sampling matter. The mean pilot validity based upon all available data is 0.20, which is slightly higher than the expected validity of 0.16 (see ch. 28) predicted on the basis of the factorial composition of the Thurstone test.

Parts I and II correlate highly with parts III and IV but not as highly as with each other. There is apparently some slight difference in the ability to select a right hand and to select a right leg, eye, or arm. Here may be evidence of restricted subfactors.

The chief interest in this test is in its factor content. Sixty percent of Hand's known common-factor variance was found on a single factor. Not more than 8 percent of the remaining common variance accumulated on any other single factor. If it is truly a relatively pure measure of a factor hitherto unknown, it is of great value. The loading in that factor should be improved if possible. If the pilot validity of the Thurstone version is actually 0.26, the margin between this and the expected 0.16 means either that the space II factor is more valid than was assumed (0.05) or there is unknown valid variance in the test.

Evaluation of Positional Discrimination Tests

Not a great deal more is known about these tests at the time of this writing than was already known concerning their Thurstone predecessors at the time work was begun on the revised forms. Validities on the new forms are generally in line with those reported earlier for Thurstone's original tests. The unique element that tests in this area have to offer is the space II factor. This factor probably has a very low validity for pilots, but a good test of it is needed. Its validity for navigators is unknown. Further study is recommended to define the factor more clearly and to maximize its variance in some test.

EVALUATION OF SPATIAL TESTS

Status of the Area

The accumulated data on tests in this area indicate that the principal unique functions measured can be explained by two factors, herein labeled space I and space II. Space I apparently has some kinship to Thurstone's spatial factor (factor S), although it is better defined. Only a select group of the tests found on factor S appear on space I. Hands, for example, originally appearing with its principal loading on Thurstone's space factor, split away from the space tests on a factor of its own (space II). Another factor can be isolated by further rotation of Thurstone's original published loadings which resembles the visualization factor described in chapter 13. Substantial visualization variance was apparently contained in factor S.

The pruning effects point out the need of further analysis, and a renaming of factor S. Thurstone's description, "facility in spatial and visual imagery," now seemingly fits the present visualization factor better than it does space I. Attempts to explain the basic nature of the factor have already been related in this chapter under the heading, Evaluation of Directional Discrimination Tests. The problems in describing and naming space II were outlined in the Evaluation of Positional Discrimination Tests, the subarea just preceding this chapter evaluation.

New Research Indicated

Further exploration of visual-spatial tests is needed along factorial lines, for apparently it is only by the application of this technique that useful conclusions can be reached. Factors cannot be well defined until adequate tests are available.

Another line of research suggested by the statistical results on tests in this area and in the visualization area reported in chapter 12 is to study the effect of the difficulty level of the items on factor composition of the test. It has been especially difficult, for example, to construct items to measure visualization that do not involve a degree of reasoning. If a visualization problem is made too difficult, it is likely solved by reasoning. By reducing the difficulty, reasoning variance seems to be reduced. If the problems are made too easy, however, they can be solved without visualization possibly by space I ability. In the Visualization of Maneuvers test, an opportunity to observe the effect of complexity is afforded. Correlations with reasoning and spatial tests are available for all three forms. The forms involving the more complex items correlate more highly with tests known to measure visualization, while the form with the simpler, more speeded items correlates more highly with tests known to measure space I. Factor-analysis results combined with systematic control of difficulty are necessary to corroborate these evidences.

BIBLIOGRAPHY

(1) Kelley, T. L. *Crossroads in the Mind of Man*, Stanford University Press, 1928.

(2) Thurstone, L. L. *Primary Mental Abilities*, Psychometric Monograph No. 1, University of Chicago Press, 1938.

CHAPTER TWENTY

Orientation Tests[1]

INTRODUCTION

Among the traits considered essential for effective air-crew performance are the abilities to determine one's bearings with respect to points of the compass and to maintain an appreciation of one's location relative to landmarks in the environment. An attempt was made, by the use of orientation tests, to measure these abilities. When construction of the orientation tests was begun, few instruments that measured orientation of any type existed. The few orientation-test items available were considered definitely inadequate.

Job-Analysis Information

In table 1.5 it may be seen that in two samples of 1,000 each of students eliminated from elementary pilot training, orientation was mentioned 13 percent and 15 percent of the time as a cause for elimination. For 100 cases of elimination from advanced single-engine training, however, this category was mentioned only 6 percent of the time; and for 100 cases of elimination from advanced twin-engine training, only 9 percent. In one sample of 100 reclassifications in operational training, orientation was mentioned only 2 percent of the time, and in another sample of the same size, it was not mentioned at all. One inference from these figures could be that orientation is a much more important factor in early pilot training than in later stages. Another could be that training eliminates relatively early the men who are deficient in this respect. There is probably some truth in both interpretations. Defects in orientation show up conspicuously in the student's failure to execute maneuvers properly and in his getting lost.

Under combat conditions, the importance of orientation ability is probably much greater than it is in training. Reference to table 1.6 will show that supervising officers rated the importance of observation and orientation for pilots as fairly high. On a 9-point scale, in which 5 means better than average, the mean ratings were 7.2 for fighter pilots and 5.5 for bomber pilots. For bombardiers and navigators in combat (see tables 1.2 and 1.4), supervisors rated the trait of orientation and observation highest in the list of 20 traits with a mean rating of 7.8 for both of these air-crew positions. Since the trait on the rating scales is

[1] Written by Capt. John I. Lacey and Sgt. S. W. Nickson.

named orientation and observation, however, it is not clear how much of this rating was based on orientation and how much on observation.

Two Types of Orientation Tests

The subareas adopted for orientation measurement are (1) compass orientation and (2) pattern orientation. This a priori division is based on the superficial requirements of the various orientation tests. Under the rubric of compass orientation are placed those tests that require the examinee explicity to use the points of the compass. The tests are Directional Orientation, Following Oral Directions, Compass Directions, and Compass Orientation. Spatial Orientation, Aerial Landmarks, and Star Identification are considered pattern-orientation tests. These tests require the examinee to identify geographical parts within a whole.

COMPASS ORIENTATION TESTS

Directional Orientation, CP515B and C [2]

Air-crew members must be familiar with the points on the compass and must also be able to apprehend directions quickly and accurately despite various conditions conducive to disorientation. Directional Orientation tests were designed to measure the speed with which directions can be accurately recognized despite various degrees of rotation of the compass rose out of its usual position (as conventionally represented). It was thought that a test of this ability is analogous to a direct test of the ability of a pilot to remain directionally oriented in spite of sudden and frequent changes in direction of flight.

Description.—In Form B, each item consists of six circles, with one direction indicated on each. Each circle is rotated out of its conventional position on the page, i. e., with north at the top and east to the right. The first circle in each problem is called the "given circle." The task of the examinee is to determine which circles of the remaining five, if superimposed on the given circle, would have indicated directions which would be in proper relationship to the indicated direction on the given circle. Thus, in figure 20.1, with N pointing as it does, circles 1, 3, and 4 match the given circle, while 2 and 5 do not.

The task of the examinee in form C is essentially the same, but the form of the problems is somewhat changed. The examinee is presented with five circles, each with but one direction indicated, and each rotated out of position as in Form B. Through each circle is drawn a diameter, at one end of which is an arrowhead. The examinee must determine which of these circles contain arrowheads pointed in a given direction. Thus, in figure 20.2, the arrows in 1, 3, and 4 point northeast, the given direction, but those in 2 and 5 do not.

[2] Developed at Office of the Surgeon, Headquarters, AAF Training Command. Chief contributors: Maj. James J. Gibson and Maj. George F. J. Lehner. These two forms and form A were based upon Dr. Paul Woodring's research on directional orientation, the manuscript of which was generously made available to the Army Aviation Psychology Program.

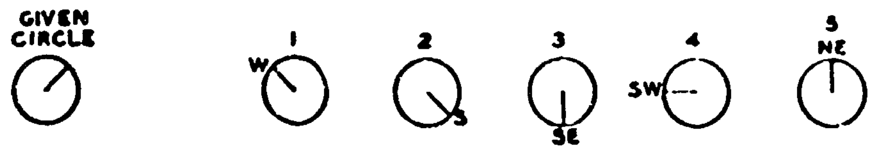


FIGURE 20.1
PRACTICE PROBLEM OF DIRECTIONAL ORIENTATION,
CP515 B

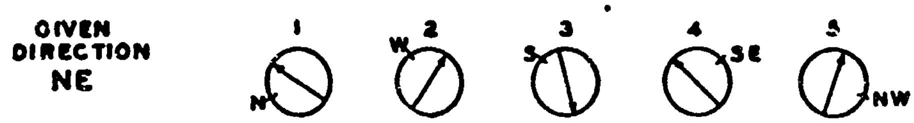


FIGURE 20.2
PRACTICE PROBLEM OF DIRECTIONAL ORIENTATION,
CP515C

(1) *Internal characteristics.*—There are 2 unscored sample items and 28 scored items in each form.

(2) *Administration.*—Reading of the directions for each form requires from 3 to 5 minutes, while testing time is 18 minutes for Form B and 20 minutes for Form C, allowing approximately 50 percent of the examinees to finish.

(3) *Scoring.*—The scoring formula is R−W. The maximum score for Form B is 70, and for Form C, 60. While separate scores for Forms B and C were obtained, they were combined in computing some of the statistics reported.

Statistical results.—Except where noted below, the following data are for examinees tested at Psychological Research Unit No. 3 in March 1943.

(1) *Distribution statistics.*—A sample of 392 unclassified aviation students yielded a mean score of 47.0 and a standard deviation of 17.4 for Form B, and a mean of 42.6 and a standard deviation of 13.9 for Form C.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.49, a standard deviation of the phi distribution of 0.21, and a range of values from 0.14 to 0.85. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent in total score (the two forms combined) of a group of 360 unclassified aviation students.

(3) *Reliability coefficient.*—Based upon a sample of 392 unclassified aviation students, a correlation of 0.74 was found between forms B and C. This figure provides a conservative estimate of the reliability of either form.

(4) *Correlation between rights and wrongs.*—For a sample of 339 pilots tested in July and August 1944 at Psychological Research Unit

No. 3, the correlation between rights and wrongs was −0.48. A correlation of −0.36 was secured for a sample of 751 navigators tested in February and March 1944, at Selman and Ellington Fields and at Psychological Research Unit No. 3.

(5) *Difficulty.*—Based upon item analysis of the responses of 360 unclassified aviation students, the test (forms B and C combined) yielded a mean proportion of correct responses of 0.68, corrected for chance, with a range from 0.33 to 0.94 and a standard deviation of 0.17.

(6) *Factorial composition.*—The noteworthy loadings of form B are in the spatial-relations (0.41), visual-memory (0.36), general-reasoning (0.31), visualization (0.26), and numerical (0.22) factors. The communality is 0.56. For a fuller picture of the factorial composition of this test, see appendix B. Form C was not analyzed because the correlational patterns indicated that the two tests were almost identical factorially.

(7) *Test validity.*—Validation results based on several samples are given in table 20.1.

TABLE 20.1.—*Validity data for Directional Orientation, CP515B, based upon the graduation-elimination criterion*

Group	Score	N_t	p_s	M_s	M_t	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training[2]	Rights	339	0.82	60.63	56.43	11.23	0.21	0.35
Do[2]	Wrongs	339	.82	4.41	4.36	4.29	.01	−.16
Do[2]	R−W	339	.82	56.22	52.07	13.86	.17	.33
Do[2]	R−W/4	339	.82	59.53	55.34	11.79	.20	.35
Pilots through basic training[3]	R−W	563	.68	38.33	31.18	17.85	.34	.31
Navigators[4]	Rights	751	.91	58.52	51.76	12.31	.28	.38
Do[4]	Wrongs	751	.91	2.79	4.28	4.08	−.19	−.27

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Tested July 6 to Aug. 12, 1944 at Psychological Research Unit No. 3.
[3] In classes 44B and 44C. Tested at Psychological Research Unit No. 3.
[4] Tested Feb. 11, 1944 at Selman Field, Feb. 1 and 2, 1944 at Elington Field, and Mar. 21, 1944 at Psychological Research Unit No. 3.

Evaluation.—Forms B and C of Directional Orientation possess a fairly high degree of homogeneity, while the items are not particularly difficult. As portrayed by factor analysis, 56 percent of the total variance of form B has been accounted for by common factors. Of the total variance, spatial-relations contributes 17 percent, visual-memory 13 percent, general-reasoning 10 percent, visualization 7 percent, and the numerical factor 5 percent. Most of the remaining 4 percent of the total variance is accounted for by three other factors, none contributing more than 1 percent.

Against the pilot criterion, the test has very satisfactory validity; and it has a higher biserial with the navigation criterion. Its validity for the pilot can be attributed to its variances in spatial-relations, visualization, and visual memory. Its validity for the navigator is due to its variances in spatial-relations, general-reasoning, visualization, and numerical factors. No compass-orientation factor as such has appeared, but this may be because there was never more than one test involving compass directions in any analyzed battery.

Directional Orientation, CP515A [8]

Variation.—CP515A, the first Directional Orientation test constructed, has the same purpose as forms B and C, but its administrative directions and item construction are considerably different. Each test item consists of a 2 3/16 inch circle with an arrow through the circumference indicating north. Within each circle are two arrows, labeled B and C, which represent the two legs of the flight of a plane, one the path of a plane before it turns, the other, the path after it turns. It is the examinee's task to give (A) the direction he is going before entering the turn, and (B) the direction he is going after making the turn. The directions are to be given in relation to north as indicated on the circumference of the circle. The answers are recorded by filling a space under either N, NE, E, SE, S, SW, W, or NW. There are 36 items in the test, the first 6 of which are practice items. Thus, the maximum number of correct responses is 60. The scoring formula is R−W. Figure 20.3 shows three items of the test.

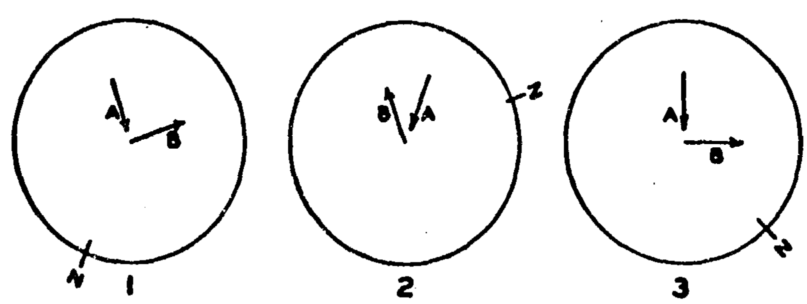


FIGURE 20.3
SAMPLE ITEMS OF DIRECTIONAL ORIENTATION, CP515A

Statistical results. (1) *Distribution statistics.*—A sample of 392 unclassified aviation students, tested in March 1943 at Psychological Research Unit No. 3, yielded a mean score of 47.0, a standard deviation of 17.4. The distribution curve was approximately symmetrical and somewhat flatter than normal.

(2) *Test validity.*—Validation results based on several samples are given in table 20.2.

TABLE 20.2.—*Validity data for Directional Orientation, CP515A, using the graduation-elimination criterion*

Group	N_t	p_g	M_g	M_e	SD_t	r_{bis}
Pilots in primary training[1]	592	0.78	37.62	31.14	17.04	0.21
Pilots through basic training[1]	563	.67	38.33	31.18	17.85	.24
Flexible gunners in training	349	.95	43.02	34.47	15.23	.26

[1] Same sample followed through the first two phases of training. In classes 44B and 44C, tested at Psychological Research Unit No. 3 in April 1943.

[8] Developed at Office of the Surgeon, Headquarters, AAF Training Command. Chief contributors: Maj. James J. Gibson and Maj. George F. J. Lehner.

Evaluation.—The distribution statistics indicate that the items in the test are of moderate difficulty and have a reasonable amount of spread in difficulty. The test appears to be moderately valid for both pilots and flexible gunners.

Directional Orientation, CP515D, E, and F [4]

The rationale for the construction of forms D, E, and F of Directional Orientation is basically the same as that for forms A, B, and C. Aerial photographs are employed in these forms, however, while diagrams are used in forms A, B, and C. The use of photographs allows not only for rotation of compass directions, but also for presenting views at various angles, e. g., vertical and oblique.

Description.—The items are composed of circular parts of aerial photographs, 2 3/16 inches in diameter. Both photographs in a pair are of the same portion of the landscape, but the second one is rotated. A single compass direction is indicated in the first photograph, and an arrow is drawn showing an unnamed direction in the second. The examinee's task is to determine the compass direction of the arrow. In forms D and F, vertical views are used for both photographs; while in form E, the initial photograph is a vertical view and the second photograph is an oblique (and rotated) view of the same terrain. Forms D and E are printed together in one booklet as parts I and II. Form F is a reprint of form D, designed for administration in a special intercorrelational study.

(1) *Internal characteristics.*—Form D includes 1 recorded but unscored sample item and 44 scored items. Form E contains 3 recorded but unscored practice items and 45 scored items. Figure 20.4 shows two items of form D, and figure 20.5 shows two of form E.

(2) *Administration.*—Answers are recorded directly on a special IBM answer sheet with marking spaces after each item number labeled N S E W, standing for north, south, east, and west. Eight minutes are allowed for part I, 16 minutes for part II, and 6 minutes are considered adequate for administration of directions. The total testing time, therefore, is 30 minutes.

(3) *Scoring.*—The scoring formula is $R-W/3$.

Statistical results. (1) *Distribution statistics.*—For a sample of 500 unclassified aviation students tested in July and August 1944, at Psychological Research Unit No. 3, and in March 1945, at Psychological Research Unit No. 2, the mean scores (rights only) were 29.5 and 19.9 for Forms D and E respectively; the standard deviations were 8.7 and 6.3.

(2) *Reliability coefficient.*—No satisfactory estimates of reliability are available. The correlation between Forms D and E, however, for the above-mentioned sample of 500 unclassified students, was 0.36. Since the forms are different, this is a gross underestimate of the reliability.

[4] Developed at Psychological Research Unit No. 3. Chief contributors: Sgt. Hyman Heller and Pfc. Charles W. Nelson.

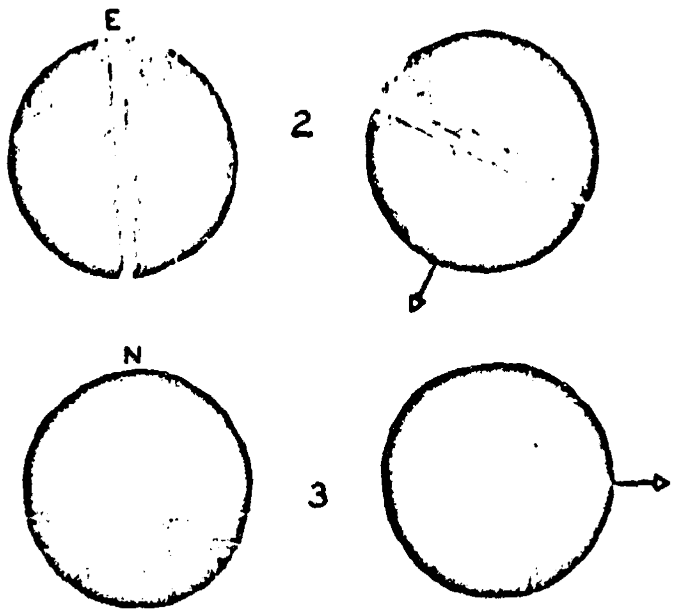


FIGURE 20.4
SAMPLE ITEMS OF DIRECTIONAL ORIENTATION,
CP515D

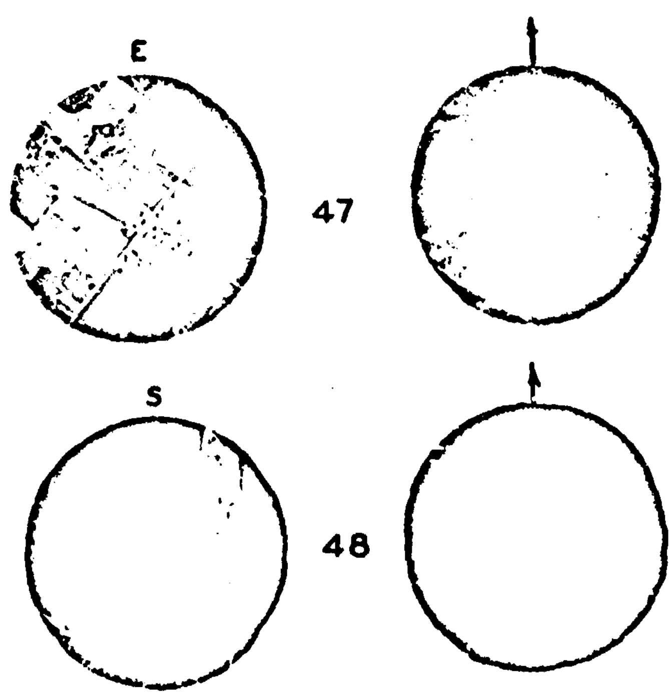


FIGURE 20.5
SAMPLE ITEMS OF DIRECTIONAL ORIENTATION,
CP515E

Evaluation.—The utilization of aerial photographs in these forms, while adding face validity, has probably had the effect of introducing some perceptual-speed variance into the test. Forms D and E have a very modest intercorrelation; apparently the addition of the feature of oblique views in Form E has introduced new factors, quite possibly held in common with Aerial Landmarks, CP525 (see pp. 535f.).

Following Oral Directions, CI651AX [5]

When this test was designed, two rationales were advanced to justify its construction. First, it was regarded as an integration test (see ch. 10). The rationale for all integration tests also applies here. Second, it was regarded as a directional-orientation test. It was considered to be especially important for the combat pilot to be able to react quickly and correctly to the movements of enemy aircraft. He must rapidly alter the course of his airplane, making continuous changes in both direction and altitude in order to out-maneuver enemy planes. All of this time he must maintain correct orientation and, in addition, maintain a constant state of readiness for surprise attacks from new positions.

Description.—Verbal descriptions are presented, by means of phonograph records, of planes flying and being attacked by enemy planes from various directions. The examinee is directed to imagine that his plane has executed a maneuver according to rules laid down in preliminary instructions. The answer to the item is the direction in which the plane is flying at the end of the maneuver. The items are made more difficult as the test progresses, by increasing the number of attacking planes from one to two, and by requiring the examinee to adhere to rules governing altitude as well as direction.

(1) *Internal characteristics.*—In part I, the problems involve one attacking plane and require an answer in terms of changed direction only. In part II, two attacking planes must be considered in sequence, but the answer again is in terms of changed direction only. Parts III and IV require answers in terms of changed direction and altitude. Part III involves one attacking plane, while part IV involves two planes attacking in sequence.

The directions contain five unscored sample items. There are 5 items in part I, presented at 4-second intervals; 10 in part II, at 8-second intervals; 5 in part III, at 6-second intervals; and 35 in part IV, at 15-second intervals. Parts I, II, and III were designed as gradual training for the complicated part IV, which was considered to be the heart of the test. This accounts for the part-differences in number of problems. The items are presented in series of five.

(2) *Administration.*—Each examinee receives a work sheet and a 15-place IBM answer sheet. Answers are marked directly on the work sheet and must be transcribed to the IBM answer sheet after the test is

[5] Developed at Psychological Research Unit No. 3. Chief contributors: S/Sgt. J. Gordon Eikin, Sgt. Nathan Kravetz, and T/Sgt. Sanford J. Mock.

701120—47—34

completed. The transcribing of answers takes approximately 10 minutes. The total administration and testing time is approximately 45 minutes.

Relevant parts of the directions to part I, which illustrate the task of the examinee, follow:

This is a test to see how well you can keep several different facts in your mind while receiving orders to change course * * *

If an enemy plane attacks from your right, you will turn to the left. If it attacks from the left, you will turn right. If you are attacked from the rear, you will continue on your course, and if you are attacked head on, you will reverse your directions.

These left, right, front, rear movements will be changed into compass directions—north, south, east, west.

For example—if you are told that you are heading north and are attacked from the left you will, according to the rule, turn right, which in this case would be east. East, then, is the new direction of your plane.

Now if you are told that you are being attacked from the west, you will continue on your course to the east, because the west in this case is the rear.

The examinee is told that, within each series of problems, * * * the direction you start from in each problem will be the direction you gave as the answer in the previous problem.

The examinees record their answers by marking N, S, E, or W after the number of the problem on the work sheet.

The sample problems of part I are:

Write the answer to this first problem on your work sheet opposite number 1.

1. You are flying west and are attacked from the left. Write on the work sheet the direction in which you are now headed. This is the new direction of your plane. You will make your next move from this direction. I shall repeat No. 1. You are flying west and are attacked from the left.
2. You are attacked from the south.
3. You are attacked from the right.
4. A plane attacks from the rear.
5. Attack comes from the left.

The item numbers are read each time before reading the item. The answers to the sample items are given, and the rules of the test are again emphasized before starting part I.

Relevant parts of the directions for part II are:

In the next series of maneuvers you will be attacked by two enemy planes in rapid succession. These planes will approach one after the other from the same or different directions. You are to change your course as the attacks are made. Remember, you change your course in responding to one plane before you consider the next. After you have evaded the first attacking plane, you are no longer concerned with it and you immediately maneuver to escape the second plane * * * your answer will be the direction in which you are headed after both attacks.

In part III the examinee is told that his actions will be affected by altitude as well as direction. The instructions administered in parts I and II still apply to part III with the following addition:

If any enemy attacks from a higher altitude than your own, you dive. If you are attacked from a lower altitude, you climb.

Practice problems are administered and explained before commencing part III.

In part IV the examinee is told that he will be subjected to attack from the same or different direction and altitude by two planes in succession and that he will complete the movement necessary to escape the first plane before responding to the second. The answer is the new direction after both enemy attacks have been evaded.

Two illustrative problems are explained before beginning part IV.

At the conclusion of part IV, answer sheets are distributed so that the responses from the work sheets may be transcribed.

(3) *Scoring.*—The scoring formula is R−W/5.

Statistical results. (1) *Distribution statistics.*—A sample of 1,302 classified pilots in class 44E (tested at Psychological Research Unit No. 3) yielded a mean score of 37.2 and a standard deviation of 9.8. The distribution curve is moderately negatively skewed.

(2) *Internal consistency.*—Based on 800 pilots tested in October 1943 at Psychological Research Unit No. 3, an item analysis was made for the highest 25 percent and the lowest 25 percent of the cases in total score. The mean internal-consistency phi is 0.40, the standard deviation 0.10, and the range is from 0.10 to 0.60.

(3) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.70, corrected for length, was obtained. This figure is based on a sample of 278 unclassified aviation students tested in September 1943 at Psychological Research Unit No. 3. Alternate forms were secured by arbitrarily dividing the test into first half and second half. Since these two halves are not entirely comparable, 0.70 is probably an underestimation of the test's reliability.

(4) *Difficulty.*—For analytical purposes, parts I, II, and III were grouped together and compared with part IV. For the first three parts, the mean proportion of correct responses, corrected for chance success, is 0.80, with a range from 0.65 to 0.97 and a standard deviation of 0.08. For part IV, the corrected figures are: Mean, 0.58; range, 0.29 to 0.77, and standard deviation, 0.12. These data are based upon the above-mentioned sample of 800 pilots.

(5) *Factorial composition.*—The strongest loadings were found in the spatial-relations (0.28), general-reasoning (0.27), integration II (0.25), numerical (0.21), and visualization (0.20) factors. The communality is 0.42. For a fuller picture of the factorial composition of this test, see appendix B.

(6) *Test validity.*—Validation results are given in table 20.3.

TABLE 20.3.—*Validity data for Following Oral Directions, C1651A, graduation-elimination criterion*

Group	N_t	r_{tt}	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Pilots in primary training[2]	1,302	0.91	37.60	34.00	9.75	0.19	0.24
Pilots through basic training[2]	1,292	.86	37.76	34.22	9.74	.20	.25

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44E. Tested at Psychological Research Unit No. 3.

Evaluation.—The adequacy of the training provided in parts I, II, and III is unknown. Part IV, however, is considerably more difficult than the preceding parts, even though the training probably aids performance in it. The simple and complex parts of this test deserve separate analysis and exploitation.

The reliability is probably satisfactory, but the pilot validity is only moderate. Only 42 percent of the total variance of this test has been accounted for by common factors. Of the total variance, the spatial-relations factor accounts for 8 percent, general reasoning 7 percent, and integration II 6 percent. The major portion of the remaining variance is well dispersed among four other factors, none contributing more than 4 percent. Contrary to what might be expected for a verbal test of this kind, the verbal factor accounts for only 3 percent of the total variance. The probable explanation is that the verbal level of the test is so low that the differences in verbal comprehension among the examinees are inconsequential. Since the communality is so much lower than the reliability, the primary future interest in this test should be in identifying its unknown variance. Since the test is exceedingly complex, it has little or no value as a classification instrument. It is possible that separate analyses of the parts would separate the factor variances somewhat. Since the pilot validity predicted from its factor loadings (0.20) is close to the obtained validity (0.24), no new factor valid for pilots is promised by the test.

Variations.—Two variations were constructed because it was thought desirable to provide two tests, one composed of the simple items of part I of the AX form, and another composed of the more complex items of part IV of that form. This decision was reinforced when it was found, for a sample of 270 cases, that the items in the first half of the test (parts I, II, III, and some items of part IV) correlated 0.26 and 0.22 with Mathematics B, C1206C, and Figure Analogies, C1212AX1; whereas the items in the last of the test (remaining items of part IV) had correlations of 0.35 and 0.38 respectively. Apparently, then, the more complex items in C1651AX have more reasoning content than the simpler items.

Following Oral Directions, CP651BX

This form is an expanded version of part I of CP651AX. There are 5 unscored practice items and 145 scored items. Four seconds are allowed for each practice item. Items 6 through 150 are administered by phonograph with 2-second intervals between items. Part I is made up of items 6 through 75, while part II consists of items 76 through 150. At the conclusion of the test, the responses are transcribed to an IBM answer sheet. Testing time, directions, and transcription require approximately 25 minutes. The scoring formula is R−W/3.

Statistical results. (1) *Distribution statistics.*—For a sample of 1,167 pilots in class 44G, tested at Psychological Research Unit No. 3, the mean score was 116.6 and the standard deviation 29.3.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.39, a standard deviation of the phi distribution of 0.09, and a range of values from 0.15 to 0.54. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent of a group of 750 unclassified aviation students tested in May 1944 at Psychological Research Unit No. 3.

(3) *Reliability coefficient.*—By the alternate-forms method, an estimated reliability coefficient of 0.85, corrected for length, was obtained, based on a sample of 487 pilots in class 44G, tested at Psychological Research Unit No. 3. For 1,682 navigators, the corrected reliability of rights is 0.82, and for wrongs, 0.77. These navigators were tested in April, May, and June 1944 at the three Psychological Research Units.

(4) *Correlation between rights and wrongs.*—Data on the correlation of rights and wrongs are given in table 20.4.

TABLE 20.4.—*Correlation between rights and wrongs for Following Oral Directions, CI651BX*

Group	N	Part	r_{rw}
Pilots in primary training[1]	734	Total	−0.75
Navigators[2]	1,682	I	−.86
Do[2]	1,682	II	−.91
Do[2]	1,682	Total	−.89

[1] Tested from May 9, 1944 to Aug. 12, 1944 at Psychological Research Unit No. 3.
[2] Tested June 1 and 2, 1944 at Psychological Research Unit No. 1; Apr. 17 to 21, 1944 at Psychological Research Unit No. 2; and May 1 to 6, 1944 at Psychological Research Unit No. 3.

(5) *Difficulty.*—Based upon a sample of 800 unclassified aviation students tested at Psychological Research Unit No. 3 in May 1944, the mean proportion of correct responses, corrected for chance success, is 0.68, with a range from 0.33 to 0.95, and a standard deviation of 0.15.

(6) *Test validity.*—Validation results are given in table 20.5.

TABLE 20.5.—*Validity data for Following Oral Directions, CI651BX*

Group	Criterion	Score	Part	N_t	p_a	M_p	M_a	SD_t	r_{bis}	$_{c}r_{bis}$[1]
Pilots in primary training[2]	Graduation-elimination	R−W/3	Total	1,167	0.90	118.51	99.30	29.31	0.19	0.22
Do[3]	....do	R−W/3	....do	1,771	.89	113.41	101.20	32.14	.20	.23
Do[3]	Ability-ratings[4]	R−W/3	....do	700	.90	118.55	105.35	29.72	.23	.26
Do[3]	Graduation-elimination	Rights	....do	734	.76	107.34	98.30	25.45	.21	.29
Do[3]	....do	Wrongs	....do	734	.76	26.53	35.21	19.99	−.26	−.32
Navigators[5]	....do	Rights	I	1,682	.92	57.51	51.39	11.45	.27	.35
		....do....	II	1,682	.92	68.76	64.87	8.85	.22	.30
		....do....	Total	1,682	.92	126.27	116.26	18.66	.27	.36
Do[6]	....do	Wrongs	I	1,682	.92	9.71	13.79	9.25	−.22	−.31
		....do....	II	1,682	.92	5.07	7.64	7.37	−.17	−.26
		....do....	Total	1,682	.92	14.78	21.43	15.04	−.22	−.31
Do[6]	....do	R−W/3	....do	1,682	.92	121.34	109.12	23.23	.26	.35

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44G, tested at Psychological Research Unit No. 3.
[3] In classes 44G and 44G, tested at Psychological Research Unit No. 3. Partially overlaps first sample.
[4] Pilot trainees routinely receive ratings as Above Average, Average, and Below Average. The dichotomy used contrasts the below-average group, including eliminees, with the average and above-average group.
[5] Tested from May 9 to Aug. 12, 1944 at Psychological Research Unit No. 3. Does not overlap previous sample.
[6] Tested June 1 and 2, 1944 at Psychological Research Unit No. 1; Apr. 17–21, 1944, at Psychological Research Unit No. 2; and May 4–6, 1944, at Psychological Research Unit No. 3.

(7) *Item validity.*—Validation of items revealed a mean phi of 0.07, with a range from −0.06 to +0.20, based upon the responses of 600 graduates and 117 eliminees from primary pilot training in class 44G, tested in December 1943 and January 1944 at Psychological Research Unit No. 2.

Evaluation.—The reliability of this test is quite satisfactory. Its validity for pilots seems to be about the same as for the AX form. Navigator validities are considerably higher. One noteworthy finding is that the rights alone are perhaps more valid for navigators than is a formula score. The wrongs have considerable dispersion and validity in their own right, and a study of optimal scoring formula seems indicated.

Following Oral Directions, CI651CX

This form includes the same type of items and administrative directions as part IV of CI651AX. There are 3 practice items and 40 test items, with 15-second intervals between test items. Administration, testing, and transcription of answers require approximately 40 minutes. The scoring formula is R−W/5. No statistical data concerning this test are available at present.

Compass Orientation, CI660A [4]

This test was developed as a result of the promising validity reports on Following Oral Directions, and it is an attempt to provide a simple and pure test of the function it was thought was unique in Following Oral Directions. This unique contribution was thought to consist of the measurement of the ability to orient rapidly to changes in directions.

Description.—In each item, one of the four compass directions, north, south, east, or west, is presented as an initial direction of flight. Then a turn, either left or right, is given. The examinee's task is to record the new direction of flight after the turn is made. The mode of presentation of the items is as follows:

Item	*You are flying*	*And turn*	*New Direction*
80.	North	left	______________
81.	West	right	______________
82.	East	left	______________

(1) *Internal characteristics.*—The instructions contain 2 items with the correct answers marked and 28 recorded but unscored practice items. There are 150 scored items in the test.

(2) *Administration.*—Each examinee receives a work booklet with directions and printed items. Answers are marked directly in the work booklet and must be transcribed to the standard IBM answer sheet when

[4] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Lloyd G. Humphreys, Pvt. James A. Walker, and Lois G. Wright.

the testing period is ended. The examinee is allowed 50 seconds to complete the 28 sample items. Five minutes are allowed for working the 150 scored items in the test. Administration consumes approximately 2 minutes, and transcribing takes about 5 minutes, making a total testing time of approximately 12 minutes. The examinee is told that the test is a speed test and that his score will be simply the number of correct responses.

(3) *Scoring.*—Rights and wrongs were scored separately.

Statistical results. (1) *Distribution statistics.*—A sample of 578 unclassified aviation students tested in October 1944 at Psychological Research Unit No. 2 yielded a mean rights score of 95.7 and a standard deviation of 32.8. The distribution curve is moderately negatively skewed and somewhat flatter than normal.

(2) *Internal consistency.*—The degree of homogeneity of the items of the test is indicated by a mean internal-consistency phi of 0.26, a standard deviation of the phi distribution of 0.15, and a range of values from −0.15 to 0.95. These statistics are based upon an analysis of the responses of the highest 27 percent and the lowest 27 percent of a group of 750 unclassified aviation students tested in July 1944 at Psychological Research Unit No. 3.

(3) *Difficulty.*—Based upon the responses of the above-mentioned sample of 750 unclassified aviation students, the test yielded a mean proportion of correct responses of 0.88, corrected for chance, with a range from 0.43 to 1.00 and a standard deviation of 0.10.

(4) *Test validity.*—Validation results based on a single sample are given in table 20.6.

TABLE 20.6.—*Validity data for Compass Orientation, C1660A, based upon a sample of pilots in elementary training, graduation-elimination criterion (N=893, p_s=.84)*[1]

Score	M_s	M_f	SD_t	r_{bis}	$_c r_{bis}$[2]
R	[illegible]	89.29	22.41	0.07	0.16
W	4.24	4.64	8.70	−.03	.05
R − W	88.05	84.65	27.13	.07	.15

[1] Tested July 9 to Oct. 7, 1944, at Psychological Research Unit No. 3.
[2] Assuming an unrestricted stanine standard deviation of 2.00.

Evaluation.—The negative skew of the distribution and the high proportion of correct responses indicate that the items in the test are relatively easy, as they should be in a speed test. The validity figures for pilots are not impressive. It is possible that a navigator or bombardier criterion would yield better results. The test has not been factor analyzed. Correlational data available, however, indicate that this test would define a new factor on which Following Oral Directions and Directional Orientation would have moderate loadings. The simplicity and ease of administration of this test are most appealing.

Compass Directions, CP524A [7]

This is a test of the examinee's ability to reorient himself to a particular ground pattern quickly and accurately when compass directions are shifted about.

Description.—For every five test items there is a schematic circular map which represents an aerial view of the ground, like that shown in figure 20.6. These diagrams include such landmarks as streams, roads, supply dumps, airports, and villages. For each item, a statement is given that establishes arbitrary directional relationships on the map. The examinee is then required to answer a question in terms of the rotated compass points. Sample problem I, for example, which is used in the directions to explain the test to the examinees, reads as follows:

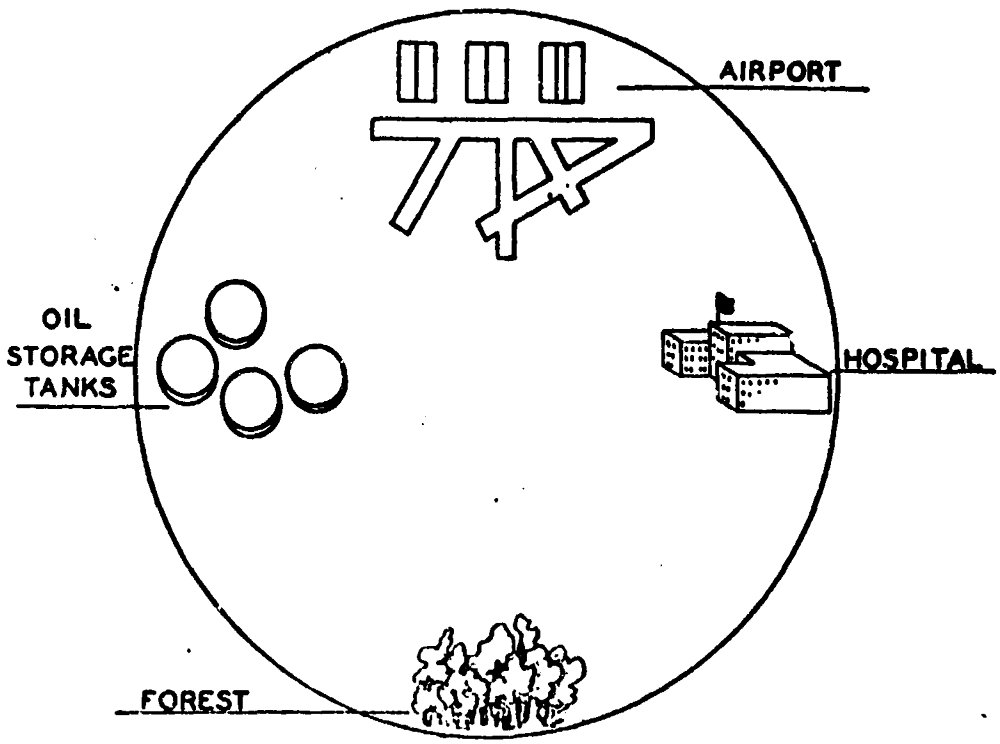


FIGURE 20.6
SAMPLE SCHEMATIC MAP OF COMPASS DIRECTIONS, CP524A

"The forest is due NORTH of the airport." (See fig. 20.6.) You are flying over the hospital. In which direction must you fly to reach the airport? (A) North, (B) southwest, (C) southeast, (D) northeast, (E) east.

The directions then continue:

The statement that the forest is north of the airport tells you that for this problem north is at the bottom of the map, south at the top, east at the left, and west at the right. The airport is to the south and to the east of the hospital. The correct answer in this case is southeast, which is choice (C).

[7] Developed at Medical and Psychological Examining Unit No. 7. Chief contributors: Cpl. Morris Freedman and Sgt. Robert Levine.

Other problems in the test call for answers in terms of compass headings in degrees. Thus, sample problem 2, also used in the directions, reads:

The oil storage tanks are east of the airport. You are taking off from the airport to bomb an enemy position in the forest. Your compass heading will be: (A) 180°, (B) 90°, (C) 270°, (D) 45°, (E) 0°.

(1) *Internal characteristics.*—There are 2 sample items (quoted above) and 30 test items, 5 to a diagram.

(2) *Administration.*—The principles involved in the test and the two sample problems are explained before beginning the test. The examinees are cautioned not to turn the test booklets in order to get north into its usual position. A testing time of 20 minutes is allowed, although in one administration of this test, the time was 25 minutes.

(3) *Scoring.*—The scoring formula is R−W/4.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 20.7. The distribution is slightly negatively skewed.

TABLE 20.7.—*Distribution constants for Compass Directions, CP524A*

Group	N	Score	M	SD
Unclassified aviation students[1]	398	R	12.6	3.9
Navigators[2]	381	R	17.1	4.9
Do[2]	781	R	15.3	4.9
Do[2]	381	W	5.1	3.2
Do[2]	381	R−W/4	15.8	5.3

[1] Tested Mar. 6-8, 1944 at Medical and Psychological Examining Unit No. 7, with a 25-minute test period.
[2] Tested May 22 and 23, 1944 at Psychological Research Unit No. 2, with a 20-minute test period.

(2) *Test validity.*—Validation results based on a single sample are given in table 20.8.

TABLE 20.8.—*Validity data for Compass Directions, CP524A, based upon a sample of 381 navigators (p_s=0.93; graduation-elimination criterion)*[1]

Score	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]
R	17.43	12.88	4.90	0.44	0.63
W	5.04	6.16	3.18	−.17	−.24
R−W/4	16.17	11.34	5.32	.43	.62

[1] Tested May 22 and 23, 1944 at Psychological Research Unit No. 2.
[2] Assuming an unrestricted stanine standard deviation of 2.00.

Evaluation.—This test has considerable promise as a satisfactory predictor of navigator success. The validity for the navigator training criterion (0.63) is one of the highest test validity coefficients reported in the Aviation Psychology Program. The test is probably quite reliable, and the difficulty level appears to be satisfactory. It has a feature not common to those discussed before in that it requires the examinee to shift orientation successively to *one* stimulus background. How important this is, and what variance it introduces, is unknown. The problem deserves serious study.

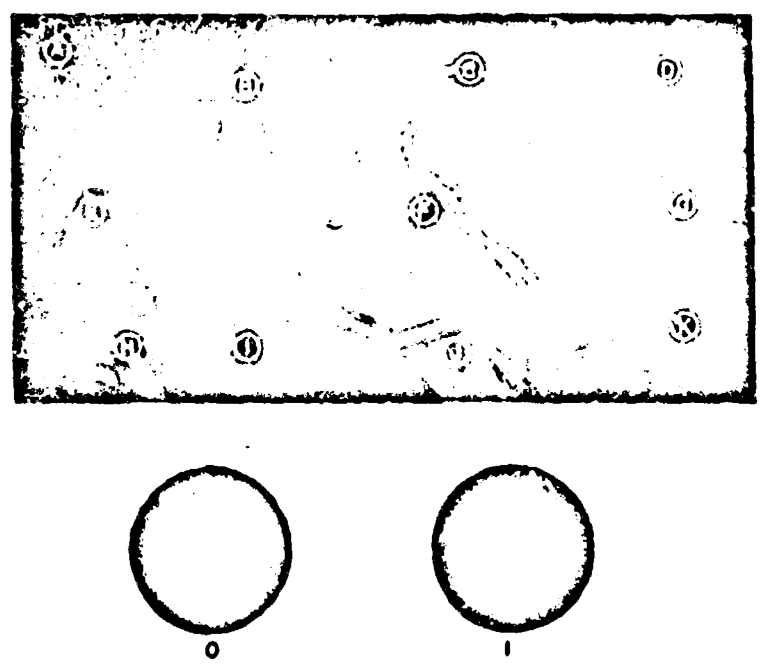


FIGURE 20.7
SAMPLE PROBLEM OF SPATIAL ORIENTATION,
CP501B

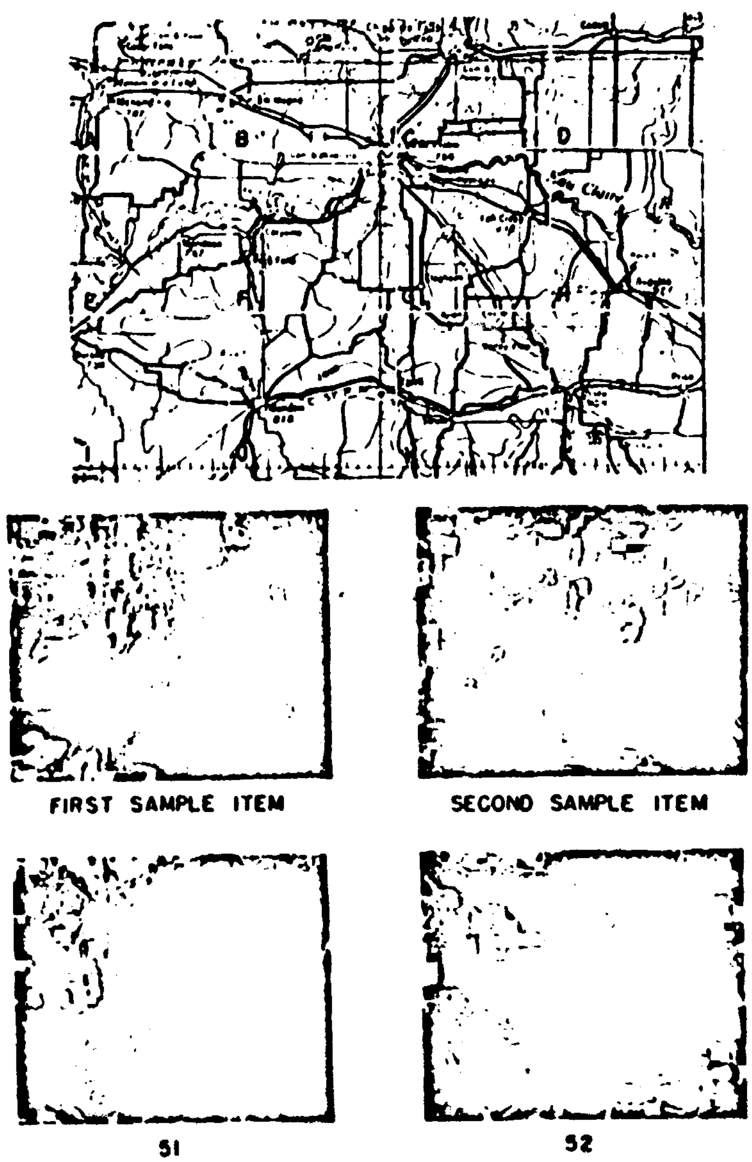


FIGURE 20.8
SAMPLE ITEMS OF SPATIAL ORIENTATION,
CP503B

Spatial Orientation, CP501B–CP503B [a]

The success of an aerial mission is frequently dependent upon the ability of air-crew members to identify points or areas on the ground which have been depicted to them in photographs or maps. Tests CP501B and CP503B were designed to measure this ability. Various forms of both these tests have been in the classification battery since August 1942.

Description.—The two tests are referred to as part I and part II respectively, since they appear in the same booklet. For part I there is a large aerial photograph at the top of each page, while directly below it are six circular photographs, 1 9/16 inches in diameter, which are parts of the larger one above. At various points in the large photographs are letters A through M. Each of the smaller photographs has a number below it that designates the number of the item. In performing the test, the examinee scans the large photograph and finds the area that matches the small photograph. The letter in the large photograph nearest the selected area is the answer to the item. The answer is recorded on a 15-place IBM answer sheet.

In part II there is a section of a standard aviation map in color on each page. Each map is sectioned off into twelve squares labeled A through L consecutively. Below the map are four 3 x 2½-inch aerial photographs of portions of the area portrayed in the map. (The scale of the photographs is 10 times that of the map.) The answer to an item is the letter of the square on the map containing the photographed area.

(1) *Internal characteristics.*—Part I contains 1 sample problem and 49 scored items based on 9 large aerial photographs. Figure 20.7 is a sample item of part I.

Part II contains 2 sample items and 50 scored items, based on 13 aerial maps. There are six sections in part II, each composed of a double page. Each page contains an aerial map and four aerial photographs to go with it. Figure 20.8 is an illustration of a sample problem in part II.

(2) *Administration.*—In the administration of part I, the task and sample item are explained. Item 1 is worked, recorded, and explained before the examinees start the test. Five minutes testing time is allowed for part I, and administration takes approximately 5 minutes.

In part II the task is described, the conventions of map representations are explained, and the two sample items are explained in detail. Three minutes' working time is allowed for each double-page section, at the end of which the examinee is told to turn the page and continue with the next item. The testing time is 18 minutes, while explanations and directions require approximately 12 minutes, bringing the total time to 30 minutes.

[a] Both forms developed at Office of the Air Surgeon, Headquarters, AAF. Chief contributors: Capt. Launor F. Carter, Capt. John T. Dailey, and Lt. Frank J. Dudek.

(3) *Scoring.*—R−W/5 was the usual scoring formula used, but on occasion, for convenience, 2R — 2W/5 was used. After August 1944, the scoring formula for parts I and II was R−W+20. Very late in the program it was found that validity-maximizing formulas were R−3.3W for form CP501B, and R — 1.3W for CP503B. These formulas were based on the statistical results yielded by a sample of 3,055 classified pilots in class 43H. The effect of utilizing this selected sample is unknown.

Statistical results for Spatial Orientation, CP501B. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 20.9.

TABLE 20.9.—*Distribution constants for Spatial Orientation, CP501B*

Group	Score	N	M	SD
Unclassified aviation students[1]	R−W/5	1,850	27.2	5.7
Do[2]	R−W/5	1,920	28.3	5.6
Do[3]	R−W/5	3,000	27.8	5.6
Do[4]	2 (R−W/5)	1,096	56.9	11.3
Classified pilots[5]	R−W+20	3,151	57.3	11.0
Armorers in training[6]	2 (R−W/5)	376	50.7	11.9
Aviation mechanics in training[7]	2 (R−W/5)	232	50.9	10.4

[1] Tested in October 1942 at Psychological Research Unit No. 1.
[2] Tested with the November 1943 Classification Battery at Medical and Psychological Examining Unit No. 10.
[3] Tested in July 1943 at Psychological Research Units Nos. 1, 2, and 3.
[4] Tested with the December 1942 Classification Battery at Psychological Research Unit No. 1.
[5] In class 43J. Tested at Psychological Research Units Nos. 1, 2, and 3.
[6] Tested with the December 1942 Classification Battery. In training at Lowry Field.
[7] Tested with the December 1942 Classification Battery at Psychological Research Unit No. 2. Entered training at Sheppard Field in March to July 1943.

(2) *Reliability coefficients.*—Three samples yielded the estimates of reliability given in table 20.10.

TABLE 20.10.—*Estimated reliability coefficients for Spatial Orientation, CP501B, based upon samples of unclassified aviation students*

N	Type	r'_{tt}	r_{tt}
183	Test-retest[1]	....	0.69
712	....Do[2]	....	.63
1,000	Part I–Part II[3]	0.84	.63

[1] Retest after 12 days, although for a few it was 6 days and for a few others 15 days. Tested at Psychological Research Unit No. 2. Data reported April 1943.
[2] Retest after approximately 30 days. At Medical and Psychological Examining Unit No. 6 from Apr. 11–19, 1945.
[3] Test divided for experimental purposes into separately timed halves. Tested at Medical and Psychological Examining Unit No. 8. Data reported May 1945.

(3) *Factorial composition, CP501B.*—The only important loading of this test is in the perceptual-speed factor (0.62). The next highest loading is in the psychomotor-speed factor (0.21). The communality is 0.69, so all the nonerror variance of the test is known. For a fuller picture of the factorial composition of the test, see appendix B.

(4) *Test validity.*—Typical validation results based on several samples are given in tables 20.11 and 20.12.

Evaluation.—With the test-retest technique, a satisfactory but moderate reliability was found.

TABLE 20.11.—*Validity data for Spatial Orientation, CP501B, using graduation-elimination criteria*

Group	Psychological Research Unit No.	Class	Score	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_s r_{bis}$[1]
Pilots in primary training	2	43D	R−W/5	1,520	0.75	27.9	25.1	5.5	0.30	...
Do	2	43E	R−W/5	1,148	.76	27.2	25.6	5.4	.17	...
Do	2	43F	R−W/5	2,375	.63	27.9	26.3	5.6	.18	...
Pilots in basic training	2	43F	R−W/5	1,428	.84	28.1	27.1	5.5	.10	...
Pilots in primary training	1, 2, 3	44E	R−W/5	4,779	.88	28.70	27.16	5.40	.15	0.18
Do	1, 2, 3	44J	R−W/5	3,145	.84	28.67	27.78	5.89	.09	.17
Do	1, 2, 3	43J	R−W+20	3,151	.66	57.3	54.2	11.0	.17	.22
Pilots through advanced training	1, 2, 3	43J	R−W+20	2,978	.54	57.59	54.59	10.97	.18	.23
Pilots in B-17 transition training	1, 2, 3	43J, K	R−W+20	1,046	.98	57.2	54.5	11.1	.10	...
Pilots in B-24 transition training	1, 2, 3	43J, K	R−W+20	983	.92	57.6	54.7	11.1	.13	...
Pilots in B-25 transition training	1, 2, 3	43J, K	R−W+20	313	.90	57.1	61.7	10.0	−.30	...
Pilots in B-26 transition training	1, 2, 3	43J, K	R−W+20	380	.82	57.9	56.0	10.8	.10	...
Pilots in P-40 assignment	1, 2	43J, K	R−W+20	2,416	.64	58.2	57.9	10.7	.02	...
Navigators	1, 2	43-10, 11	R−W+20	731	.87	59.6	51.5	11.4	.38	.41
Do	1, 2, 3	43-12 to 43-15	R−W+20	1,969	.79	58.2	53.6	11.4	.23	.31
Bombardiers[2]	1	43-5, 6, 7	R−W/5	552	.84	24.2	23.3	5.5	.09	...
Do[2]	2	43-5, 6, 7	R−W/5	330	.86	28.3	24.8	5.7	.33	...
Do[2]	3	43-5, 6, 7	R−W/5	496	.82	27.6	26.4	5.9	.11	...
Do[3]	1, 2, 3	43-8 thru 43-11	R−W+20	1,829	.79	58.06	56.16	10.93	.10	.13
Do[3]	1, 2, 3	43-14 thru 43-18	R−W+20	456	.84	57.5	56.6	11.0	.04	.08
Do[4]	1, 2, 3	43-14 thru 43-18	R−W+20	524	.86	63.3	64.8	9.9	−.08	.00
Radio operator-mechanics	1, 2, 3	....	R−W+20	235	.65	54.49	54.48	10.62	.00	...
WASPs[5]	....	44-W-7	R−W/5	91	.80	31.16	26.11	7.22	.40	...
Do[5]	....	44-W-8	R−W/5	104	.61	29.48	25.73	6.39	.36	...

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] New aviation cadets taking the 12-week course (no navigation training).
[3] New aviation cadets taking the 18-week course (some navigation training).
[4] Reclassified pilots taking the 18-week course.
[5] Women's Army Service Pilots.

TABLE 20.12.—*Validity data for Spatial Orientation, CP501B, using various criteria*

Group	Psychological Research Unit No.	Criterion	N	r[1]	$r_{corr.}$[2]
Air mechanics in training	1, 2, 3	Average grades	232	0.24	...
Armorers in training	1, 2, 3	....do	376	.07	...
Radio operator mechanics in training	1, 2, 3	....do	153	.10	...
Pilots in training, class 43J	1	Fixed gunnery, total percent hits[3]	459	.12	...
Pilots in training, class 43K	1	....do[3]	403	.05	...
Pilots in advanced single engine training, classes 43J to 44A	3	Proficiency rating	562	.12	...
Pilots in advanced twin-engine training, classes 43J–44A	3	....do	685	.05	...
Bombardiers, classes 43-1 to 43-4	3	Average grades	195	.07	...
Do	3	Record circular error[3]	195	−.10	...
Navigators in Hondo classes 43-10 through 43-18	1, 2, 3	Grades in dead reckoning (ground)	463	.14	.22
Do	1, 2, 3	Grades in celestial navigation (ground)	463	.14	.21
Do	1, 2, 3	Grades in dead reckoning (flight)	463	.10	.15
Do	1, 2, 3	Grades in celestial navigation (flight)	463	.12	.18
Do	1, 2, 3	Grades in meteorology	463	.10	.17
Do	1, 2, 3	Military grades	463	−.02	.02
Do	1, 2, 3	Final composite grades	463	.16	.24

[1] Product-moment correlation.
[2] Assumed unrestricted stanine standard deviation not reported.
[3] Highly unreliable criterion.

Since this test was included in several analyses, considerable factor-analysis data are available. The weighted averages of these data show that 69 percent of the total variance of this test has been accounted for by common factors. The perceptual-speed factor contributes 38 percent, and the remaining variance is scattered over several other abilities, not exceeding 4 percent in any one factor. The test is, therefore, a relatively pure measure of perceptual speed, but it is outstripped for this purpose by Speed of Identification, CP610A, by a small margin. Since the pilot validity of this test estimated from known factor loadings is very close to its obtained validity, and since the communality almost exactly equals the reliability (test-retest estimate), it can be assumed that all of its factors valid for pilot training have been accounted for.

The validity figures for this test, based upon several criteria, are varied. The validity for pilots in primary training is significant, but not impressive. The validity is considerably higher for navigators. For the unreliable bombardier criteria, validity is exceedingly low, and for radio operator-mechanics, it is nil. Uncorrected validities for WASPs are the highest of all (r_{bis}=0.36 and 0.40), but the samples are too small to permit a conclusive interpretation.

Statistical results for Spatial Orientation, CP503B. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 20.13. The distribution curves are positively skewed and somewhat flatter than normal.

TABLE 20.13.—*Distribution constants for Spatial Orientation, CP503B (Scored R−W/5)*

Group	N	M	SD
Unclassified aviation students[1]	3,000	20.4	6.5
Do[2]	1,851	20.2	7.2
Do[3]	1,920	19.1	6.5
Armorers in training[4]	376	18.4	7.3
Aviation mechanics in training[5]	232	19.0	7.3

[1] Tested in December 1942 at Psychological Research Units Nos. 1, 2, and 3.
[2] Tested in October 1942 at Psychological Research Unit No. 1.
[3] Tested at Medical and Psychological Examining Units Nos. 4 to 10 with the November 1943 Classification Battery.
[4] Tested with the December 1942 Classification Battery. In training at Lowry Field.
[5] Tested at Psychological Research Unit No. 2 with the December 1942 Classification Battery.

(2) *Reliability coefficients.*—Two samples yielded the estimates of reliability given in table 20.14.

TABLE 20.14.—*Estimated reliability coefficients for Spatial Orientation, CP503B, based upon samples of unclassified aviation students*

N	Type	r'_{tt}	r_{tt}
712	Test-retest[1]	....	0.69
185	Do[2]	....	.69

[1] Retest after approximately 30 days. At Medical and Psychological Examining Unit No. 6 from April 11 to 19, 1945.
[2] Retest after 12 days for most of the group, although for a few it was 6 days, and for a few others 35 days. Tested at Psychological Research Unit No. 2. Data reported in April 1943.

(3) *Factorial composition.*—The most significant loading is in the perceptual-speed factor (0.54). The next highest loading is in the visualization factor (0.25), and no other loading exceeds 0.20. The communality is 0.53.

(4) *Test validity.*—Typical validation results based on several samples are given in tables 20.15 and 20.16.

Evaluation—Spatial Orientation, CP503B.—As evidenced by the mean scores, this test is fairly difficult. Its scores have a considerable degree of dispersion, thus indicating good discrimination between examinees. Reliability is satisfactory for a battery test.

The weighted averages of the factor loadings in this test indicate that 53 percent of its total variance has been accounted for by common factors. Of the total variance, the perceptual-speed factor contributes 29 percent, the visualization factor 6 percent, and several other factors not more than 2 percent each. Better tests are available for both the leading factors.

The validity of this test for pilots in primary training and for navigators is moderate. For bombardiers the validity is low, and for radio operator-mechanics, it is probably zero. The obtained pilot validity (0.26) almost coincides with that predicted from known factorial content (0.24), so there is no need to examine the test further for unknown sources of pilot validity.

Variations of Spatial Orientation Tests.—Previous to construction of forms CP501B and CP503B of Spatial Orientation, forms CP501A, CP502A, and CP503A were used in the classification battery, as three parts of a single test booklet.

The administrative directions and test items for CP501A were constructed in the same manner as those for CP501B. The A form is also a "photo-photo" matching test and is composed of 42 items. Since it is the least difficult of the three tests, it was administered first.

CP502A is a "map-map" matching test, wherein the examinee is to recognize the area of a large aerial map which is depicted by a smaller enlarged map. Administrative directions for this test are of the same type as those used in the other Spatial Orientation tests. There are 24 items in the test, and being considered second in difficulty, it was administered after CP501B.

CP503A is a "map-photo" matching test, which is made up of a small square of aerial map in each test item and a large aerial photograph for each six test items. The examinee's task is to pick out the area in the photograph represented by the map-area in the test item There are 24 items in the test, and being the most difficult, it was administered last in the series.

These three tests, as previously indicated, were administered in order of difficulty. The fact that they are arranged in reverse order in the test

Table 20.15.—*Validity data for Spatial Orientation, CP503B, (Scored R−W/5), using graduation-elimination criterion*

Group	Psychological Research Unit No.	Class	N_t	p_s	M_s	M_e	SD_t	r_{bis}	$_c r_{bis}$ [1]
Pilots in primary training	2	43D	1,520	0.75	21.8	18.6	6.8	0.29	.....
Do	2	43E	1,148	.76	21.1	18.5	6.7	.22	.....
Do	2	43F	2,376	.63	21.4	18.4	6.9	.26	.....
Pilots in basic training	2	43F	1,429	.84	21.6	20.2	6.7	.11	.....
Pilots in primary training	1, 2, 3	43J	3,151	.66	21.74	19.46	6.73	.21	0.27
Pilots through advanced training	1, 2, 3	43J	2,978	.54	21.96	19.73	6.73	.21	.27
Pilots in primary training	1, 2, 3	44E	4,779	.88	21.62	18.63	6.34	.25	.27
Pilots through advanced training	3	44E	2,995	.80	22.25	20.07	6.23	.20	.24
Pilots in primary training	1, 2, 3	44I	3,145	.84	22.1	19.9	6.2	.19	.28
Pilots in B-17 transition training	1, 2, 3	43J, K	1,046	.98	21.9	20.9	6.9	.06	.....
Pilots in B-24 transition training	1, 2, 3	43J, K	982	.92	22.0	21.0	6.5	.08	.....
Pilots in B-25 transition training	1, 2, 3	43J, K	314	.98	22.8	17.1	6.6	.36	.....
Pilots in B-26 transition training	1, 2, 3	43J, K	380	.82	22.7	21.2	6.5	.12	.....
Pilots in P-40 assignment	1, 2	43J, K	2,416	.64	22.3	22.1	6.6	.02	.....
Navigators	1, 2	43-10, 11	730	.87	21.2	16.8	7.2	.33	.40
Do	1, 2, 3	43-12 to 43-15	1,970	.79	21.1[illegible]	18.76	7.13	.19	.22
Bombardiers[2]	1	43-5, 6, 7	552	.84	15.1	13.2	6.4	.16	.....
Do[3]	2	43-5, 6, 7	330	.86	21.2	17.9	6.9	.26	.....
Do[2]	3	43-5, 6, 7	496	.82	19.3	18.5	7.6	.06	.....
Do[2]	1, 2, 3	43-8 thru 43-11	1,829	.79	19.33	18.72	6.96	.05	.08
Do[3]	1, 2, 3	43-14 thru 43-18	455	.84	18.8	17.1	7.1	.13	.15
Do[4]	1, 2, 3	43-14 thru 43-18	524	.86	23.3	22.9	6.6	.03	.09
Radio operator mechanics	1, 2, 3	......	235	.65	18.20	19.02	6.99	−.07	.....
WASPs[5]	......	44-W-7	91	.80	23.[illegible]	17.00	7.27	.50	.....
Do[5]	......	44-W-8	104	.61	22.18	18.00	6.57	.39	.....

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] New aviation cadets taking 12-week course (no navigation training).
[3] New aviation cadets taking 18-week course (some navigation training).
[4] Reclassified pilots taking the 18-week course.
[5] Women's Army Service Pilots.

703320—47—35

Table 20.16.—*Validity data for Spatial Orientation, CP503B (scored R−W/5) using various criteria*

Group	Psychological Research Unit No.	Criterion	N	r[1]	$r_{corr.}$[2]
Air mechanics	1, 2, 3	Average grades	232	0.30	.....
Armorers in training	1, 2, 3	....do	376	.10	.....
Radio operator mechanics in training	1, 2, 3	....do	153	.29	.....
Pilots in training, class 43J	1	Fixed gunnery, total percent hits[3]	459	.08	.....
Do	1	....do[3]	403	.17	.....
Pilots in advanced single engine training, classes 43J–44A.	3	Proficiency rating	562	.12	.....
Pilots in advanced twin-engine training, classes 43J–44A.	3	....do	685	.04	.....
Pilots in 4-engine transition training, classes 43J through 44A.	3	....do	685	.07	.....
Bombardiers, classes 43–1 through 43–4.	3	Average grades	195	.16	.....
Bombardiers, classes 43–1 through 43–4.	3	Record circular error[3]	195	−.03	.....
Navigators in Hondo classes 43–10 through 43–15.	1, 2, 3	Grades in dead reckoning (ground)	463	.12	0.22
Do	1, 2, 3	Grades in celestial navigation (ground).	463	.13	.21
Do	1, 2, 3	Grades in dead reckoning (flight).	463	.05	.11
Do	1, 2, 3	Grades in celestial navigation (flight).	463	.12	.19
Do	1, 2, 3	Grades in meteorology	463	.14	.21
Do	1, 2, 3	Military grades	463	.07	.11
Do	1, 2, 3	Final composite grades	463	.15	.24

[1] Product-moment correlations.
[2] Assumed unrestricted stanine standard deviation not reported.
[3] Highly unreliable criterion.

booklet, however, caused administrative difficulty. The validity, distribution, and reliability data of these three tests compared favorably with the two forms that replaced them. Since the "map-map" test is unique in this set of tests, sample validity data might be of interest. For a sample of 1,282 pilots in class 43A at Kelley Field, 66 percent of whom were graduates from primary training, the uncorrected validity biserial correlation was 0.17. The mean score of graduates was 6.41, of eliminees 5.56, and the standard deviation of both groups combined was 3.15. For 185 navigators in classes 42-10 to 42-17, Southeast Training Center, 78 percent of whom were graduated, the uncorrected validity was 0.28. The mean score of graduates was 3.1; of eliminees 2.0; and the over-all standard deviation was 2.3.

Aerial Landmarks, CP525A[9]

In this test, an attempt was made to simulate the task so frequently performed by pilots, navigators, and bombardiers, of identifying from the air, landmarks previously seen in photographs taken from a different direction and height. Observations in combat and training indicated that air-crew members frequently fail in this task. Target identification, in particular, was a crucial part of the bombing mission in which weakness of this sort was brought dramatically to the supervisor's attention.

Description.—The test consists of a series of photographic presentations. Each presentation consists of two aerial photographs on opposite pages. On the left-hand page is given a vertical aerial view, called "the reconnaissance photograph," taken from approximately 10,000 feet above the ground. Certain points on this photograph are encircled and numbered, thus comprising the item numbers of the test. On the right-hand page is presented the second photograph, designated the "cockpit view." This photograph was taken from a lower altitude, 2,000 to 5,000 feet above the ground and from an oblique angle. Letters A through O appear at various points on the oblique photograph. The examinee's problem is to locate the landmarks, which were numbered in the reconnaissance photograph, in the oblique cockpit view. When the landmark has been identified, the letter nearest the cockpit view is recorded as the answer.

(1) *Internal characteristics.*—There are 11 photographic presentations as described and 55 items within the test booklet. The first photographic presentation and first five items are used as practice problems. See figure 20.9 for a sample item.

(2) *Administration.*—The examinees are told that their task is to locate landmarks in a target area. The photographic presentations are explained to them, and they are told how to select the correct answer. The examinees are then given time to do the first five items and record their responses. As they finish each practice item, they are told the correct answer for that particular item. Two minutes are allotted to complete the

[9] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. Stuart W. Cook, Tech./Sgt. Paul C. Davis, and Pfc. Charles W. Nelson.

items in each photographic presentation. Administration requires approximately 12 minutes, bringing the total testing time to approximately 32 minutes.

(3) *Scoring.*—The scoring formula is R−W/5.

Statistical results. (1) *Distribution statistics.*—A sample of 500 unclassified aviation students, tested at Psychological Research Unit No. 3 in September 1944, yielded a mean score of 24.4 and a standard deviation of 8.4.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.37, a standard deviation of the phi distribution of 0.12, and a range of values from 0.00 to 0.56. These statistics are based upon analysis of the responses of the highest 27 percent and the lowest 27 percent of a group of 750 unclassified aviation students tested at Psychological Research Unit No. 3 in September 1944.

(3) *Reliability coefficient.*—For a sample of 500 unclassified aviation students tested in September 1944 at Psychological Research Unit No. 3, a correlation of 0.70 was found between forms A and B of this test. Forms A and B are comparable as to content. This value may therefore be taken as the reliability of either form.

(4) *Difficulty.*—Based upon item analysis of the responses of the sample of 750 unclassified aviation students mentioned above, the test yielded a mean proportion of correct responses of 0.52, corrected for chance, with a range from 0.08 to 0.89 and a standard deviation of 0.24.

Evaluation.—This test has a fair degree of homogeneity and moderate reliability. Its items are reasonably difficult. Owing to the lack of adequate statistical data, no further evaluation is possible.

Variations of Aerial Landmarks.—Form CP525B is constructed in the same manner as CP525A and contains the same number of items. It is simply an alternate form or a continuation of form CP525A. It was constructed for the purposes of providing a wider selection of items and of determining reliability.

Form CP525C also contains 55 items, and it is constructed and administered in the same manner as Forms A and B. The items in Form C are those items in A and B which possess the highest degree of homogeneity as revealed in item analyses.

Star Identification, CP519B [19]

One of the prime requisites of the celestial navigator is the ability to orient himself to significant stars and star patterns. Not only must he be able to recognize stars and star patterns, but he must do so rapidly and from any position. These considerations led to the construction of a test that would measure the ability to recognize specific stars and star formations accurately and rapidly.

[19] Developed at Medical and Psychological Examining Unit No. 9. Chief contributor: Capt. Sidney M. Adams.

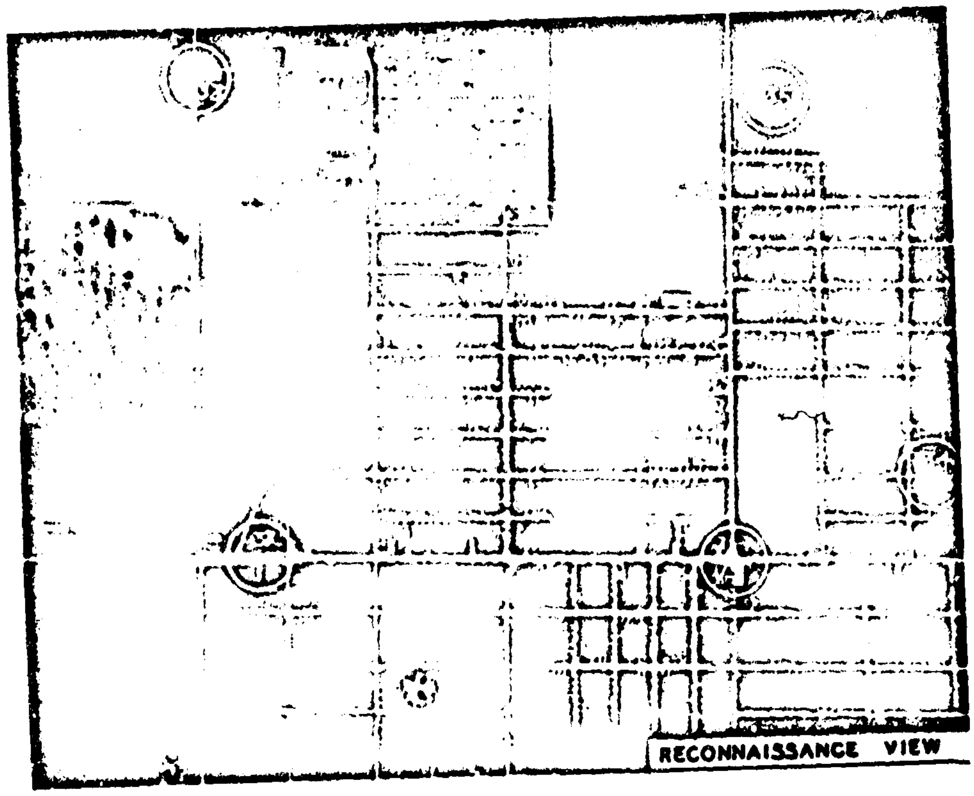
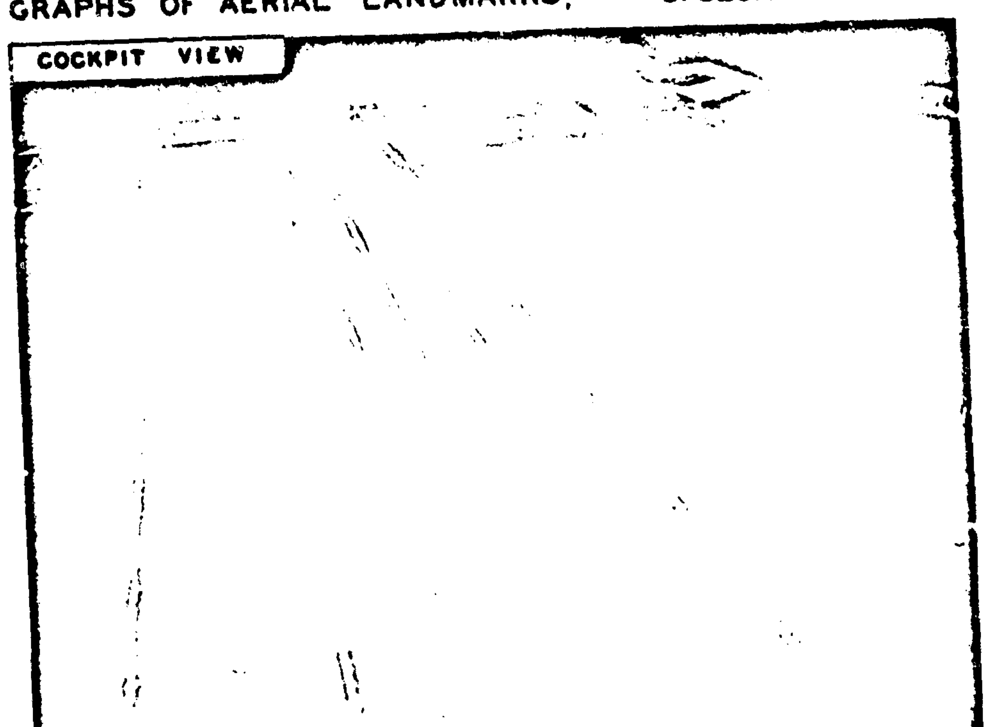


FIGURE 20.9
SAMPLE "RECONNAISSANCE" AND "COCKPIT VIEW" PHOTOGRAPHS OF AERIAL LANDMARKS, CP525A



Description.—On a separate sheet, two full-page schematic sky maps are presented. (See fig. 20.10.) For each item in the test, the examinee is required to locate a single star which is designated by number on the map. This number also establishes the item number. A test item in the booklet is made up of a square, which is an enlarged portion of a sky map, and which is divided into four parts by two intersecting lines. The four parts are labeled A, B, C, and D. Within each part are several stars, some of which, if connected with other stars in the square, would make the recognizable pattern. In doing an item, the examinee first observes

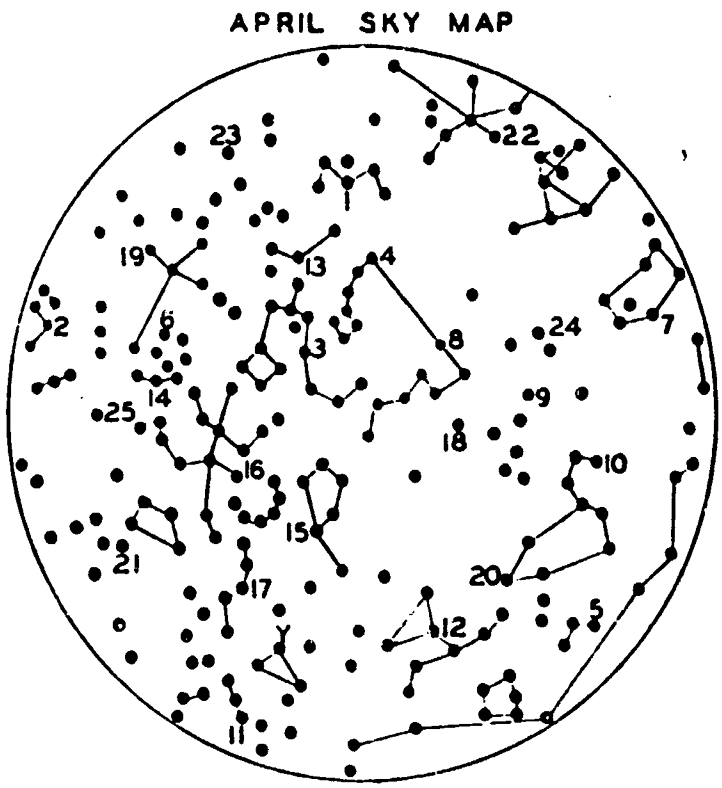


FIGURE 20.10
ONE OF THE SKY MAPS OF STAR IDENTIFICATION TEST CP519B

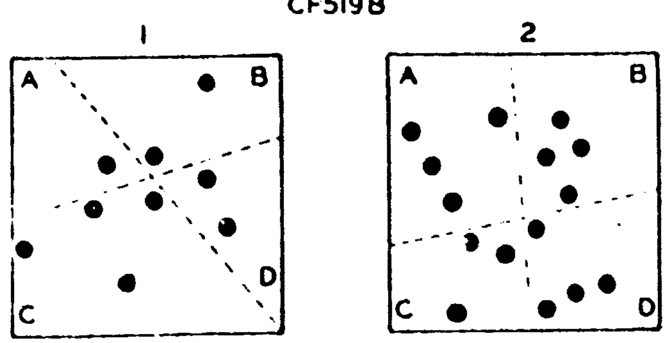


FIGURE 20.11
SAMPLE [illegible]LEMS OF STAR IDENTIFICATION, CP519 B

the numbered star on the sky map and its position in the pattern. He then solves the problem by finding the designated star in one of the four sections of the square in the booklet. Figure 20.11 shows two sample items.

Star 1 in the April sky map is in section C of the small map. Star 2 is in section B of the small map.

(1) *Internal characteristics.*—Items 1 through 25 are presented in an April sky map, while items 26 through 47 are presented in an October sky map. The booklet also contains one practice item.

(2) *Administration.*—After the practice item is explained to the examinee, he is given time to work the first two items. He is given the correct answers for these items before starting the test. Twenty-five minutes are allotted for the working of the test, while administration, as such, requires approximately 8 minutes.

(3) *Scoring.*—The scoring formula is R−W/3.

Statistical results. (1) *Distribution statistics.*—A sample of 110 pilots tested in September and October 1944 at Psychological Research Unit No. 3 yielded a mean score (rights only) of 33.0 and a standard deviation of 8.9.

(2) *Internal consistency.*—The degree of homogeneity of the items is indicated by a mean internal-consistency phi of 0.30, a standard deviation of the phi distribution of 0.15, and a range of values from 0.00 to 0.76. These statistics are based upon the responses of the highest 27 percent and the lowest 27 percent of a group of 750 classified pilots tested in October 1944 at Psychological Research Unit No. 3.

(3) *Difficulty.*—Based upon the responses of the same sample of 750 classified pilots, the test yielded a mean proportion of correct responses of 0.78, corrected for chance, with a range from 0.33 to 1.00 and a standard deviation of 0.16.

(4) *Test validity.*—Although the test was designed for selection of navigators, the only validation data yet available are for pilots (see table 20.17).

TABLE 20.17.—*Validity data for Star Identification, CP519B, based upon a sample of pilots in primary training,[1] graduation-elimination criterion*

Score	N_t	p_p	M_p	M_e	SD_t	r_{bis}	$_c r_{bis}$[2]
Rights	110	0.76	33.04	32.90	8.93	0.01	0.02
Wrongs	110	.76	6.13	6.85	4.80	−.09	−.17
R−W/3	110	.76	31.02	30.69	9.70	.04	.05

[1] Tested at Psychological Research Unit No. 3 from Sept. 13 to Oct. 1, 1944.
[2] Assuming an unrestricted stanine standard deviation of 2.00.

Evaluation.—The mean internal-consistency phi of 0.30 shows a fair degree of homogeneity, while the mean proportion of correct responses of 0.78 indicates that the test items are much too easy. This test is apparently not valid for pilot training, but if it does prove to have validity unique for navigators, it would be an excellent classification test, since it would be weighted for one assignment but not for the other. Items with a greater degree of difficulty would improve the test. The markedly superior validity of the error score in the small sample of pilots directs attention to the need for further study of the separate functions of correct and incorrect responses.

EVALUATION AND CONCLUSION

With the exception of Spatial Orientation, the tests discussed in this chapter require extensive additional research in order to evaluate them adequately. Research on most of these tests has thus far been concerned primarily with pilot validity. Since orientation tests as a class possess face validity for bombardiers and navigators, it would be important to validate such tests against bombardier and navigator criteria. More extensive factor analysis of orientation tests would supply badly needed information.

The Spatial Orientation tests I and II were used in the classification battery for over 3 years and have proved to be effective instruments, but their importance is in their measurement of the perceptual-speed factor, and not in the measurement of a primary orientation ability. Compass orientation tests have demonstrated substantial validities, particularly for pilots, without, however, exhibiting any new valid component. Results on the other orientation tests have not been carried far enough to determine whether or not they have any unique features to offer.

CHAPTER TWENTY-ONE

Tests of Set and Attention[1]

INTRODUCTION

Attention in Aviation

Job descriptions of air-crew positions stress the multiplicity of detail to which the pilot must attend in observing the dials and indicators and in operating the many switches and levers in the cockpit of a military airplane in the proper timing and sequence. This is true, although to a less extent, of the bombardier and navigator, who must also concentrate upon their tasks, often under conditions of great stress.

Job-analysis data.—Reference to chapter 1 will show how important, relatively, attention is regarded in aviation, either in training or in combat. In a study of the causes of elimination of 1,000 pilots from primary training, insufficient ability to divide attention was mentioned in 41 percent of the cases. This deficiency also existed in 43 percent of 100 cases in a study to determine the reasons for elimination from advanced training.

Of 112 instructors who were asked to indicate the causes of elimination of 232 students from navigator training, 68 percent checked "Inability to concentrate effectively over prolonged periods of time" (during examinations and flights).

Supervisors of fighter and bombardment groups indicated on a 9-point scale the minimum amount of a number of psychological traits they believed necessary for the successful completion of the pilot's combat mission. Division of attention, with the mean ratings of 7.5 and 6.8 respectively, was tied for fifth rank in a list of 20 traits for fighter pilots and was fourth from the top for bomber pilots.

Supervisors of combat bombardiers and navigators made similar ratings for these two jobs. In both cases division of attention, with a mean rating of 6.8, was in eighth place from the top rank.

Previous military studies of attention.—Most studies of attention during World War I were done by Italian psychologists. A. Gemelli (2) believed that the pilot must possess both depth of attention and ability to attend to several stimuli simultaneously. He measured depth of attention by exposing a small figure for a short period of time. The candidate was given successively longer exposures until able to perceive the figure. Records were kept of the time of exposure required for a correct

[1] Written by Staff/Sgt. Benjamin Fruchter.

response. Candidates failed this test if the time of exposure required for simple figures was greater than six-tenths of a second.

In order to measure breadth of attention, five small figures were exposed simultaneously for successively longer periods until all five (different ones in each exposure) were reported correctly. The qualifying score was 5 figures correct after one-tenth of a second exposure. No attempt was made to relate test results to flight performance.

F. U. Saffiotti (3) describes briefly, and not very clearly, an attention test that is apparently of the cancellation variety. The materials consisted of 5 series of 20 symbols each. Relevant symbols were so dispersed that a labyrinth was formed through a mass of irrelevant symbols. The subject crossed out the appropriate symbols in going through the mass. Errors and time were recorded.

Dotting tests were also used as measures of continued maximum voluntary concentration of attention.

Types of Tests to Be Considered

The tests to be discussed in this chapter were designed to measure (1) sustained attention, (2) attention under distraction, and (3) change of set.

TESTS OF SUSTAINED ATTENTION

Test of Attention, CI659AX1 [a]

This test is modeled after the two best measures of attention found by Wittenborn (4) in a factorial analysis of a battery of attention tests. In devising items for these tests, Wittenborn formulated the following requirements:

1. The performances should not depend too much upon intellectual level.
2. The tasks should depend to as small a degree as possible upon content and knowledge.
3. The tests should correlate as little as possible with factors heretofore identified.
4. The scores on the test should depend to a large degree upon a continuous, sustained application of mental effort. The tests should be so constructed that a layman might say they required a high degree of concentration.

In an attempt to devise a test that would not depend too much upon a particular kind of knowledge, material that was familiar to all persons was chosen, namely, digits and letters of the alphabet. A variety of tests was tried, and the kind of test that most dependably required continuous concentration seemed to be that patterned somewhat after tests ordinarily called following-directions tests. The items are presented one at a time at a rate that allows little opportunity for interpolated activity. The

[a] Developed at Psychological Research Unit No. 3. Chief contributors: Sgt. Roy C. Anderson, Staff/Sgt. Benjamin Fruchter.

importance of continuous and concentrated work was enhanced by using tasks in which the response to an ensuing item is in part conditioned by responses to previous items.

Description.—In part I, sets of three numbers each are presented orally by means of phonograph records. Two responses to each set are possible, depending on the interrelationships of the three numbers. The numbers in part I were recorded at the speed of 75 beats per minute. Each number is spoken on a beat with one beat skipped between sets. Precise timing was accomplished by using a metronome, which was placed behind a window in view of the speaker, but out of sound range of the microphone as the record was being cut.

In part II, words are spoken that begin with any of the first five letters of the alphabet. Words with one syllable and three letters are designated short. Words with two syllables and five letters are designated long. Two responses are possible to each word, depending upon the length of the word and the relationship of the first letter to the first letter of the preceding word.

(1) *Internal characteristics.*—Part I contains a sample series of 7 unrecorded and unscored items, and a practice series of 45 recorded and unscored items presented by phonograph, followed by 90 scored items.

Part II contains a sample list of 10 unrecorded and unscored items, a practice list of 15 recorded but unscored items, and a practice series of 45 recorded but unscored items presented by phonograph, followed by 90 scored items. Total testing time is 45 minutes for parts I and II.

(2) *Administration.*—Two types of five-place answer sheets are required for Test of Attention, CI659AX1. In part I, the answer spaces are lettered A to E. In part II, the answer spaces are numbered 1 to 5. It was felt necessary to use the numbered instead of the lettered sheet for part II to avoid confusion. Answers in part II depend upon the interrelationships of the first five letters of the alphabet. If answers had to be given in terms of letters, therefore, the task would become spuriously more difficult.

When part I is finished, the answer sheets for that part are collected before the blanks for part II are distributed. Following are the directions and sample items for parts I and II. The words in italics are spoken by the administrator and do not appear in the printed booklet of directions.

This is a test of your ability to concentrate. Numbers will be read to you from a phonograph record, three at a time. Your task is to listen to the numbers as they are read and then to blacken-in either space A or space B on your answer sheet according to the following directions which you will be given plenty of time to learn. For each set of three numbers:

Blacken space A when:

1. The first number is smallest, and the second is largest—or;
2. The first number is largest, and the third is smallest.

Blacken space B in all other cases.

Note the sample series at the right.

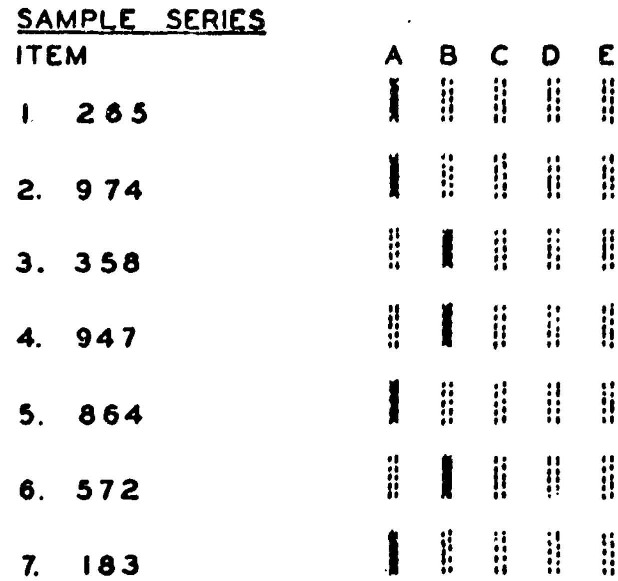


FIGURE 21.1
SAMPLE SERIES OF TEST OF ATTENTION,
C1659AX1

I will now illustrate the way each series will be introduced and read. (Pause one-half second between numbers and one second between sets.) *The next item is number 1: 2-6-5, 9-7-4, 3-5-8, 9-4-7, 8-6-4, 5-7-2, 1-8-3, and so on to the end of the series. Then the next series of 15 items will be introduced.*

In item No. 1, the first condition exists; the first number is smallest and the second largest, so the space under letter A has been blackened. (Pause) *In item No. 2, the second condition exists; the first number is largest and the third is smallest, so the space under letter A has been blackened for item No. 2.* (Pause) *In items No. 3 and No. 4, neither condition exists, so B has been blackened for these two items.* (Pause) *In item No. 5, the first number is the largest and the third number is the smallest, so A has been blackened after 5.* (Pause) *In item No. 6 neither condition exists, so B has been blackened after 6.* (Pause) *In item No. 7, the first number is the smallest and the second number is the largest, so A has been blackened after 7.* (Pause)

You will have 2 minutes in which to memorize the above rules. Go ahead. (After 2 minutes.) *Turn to page 2.*

After the practice series the instruction is, "Now we are ready to begin the test. Start with item No. 61. Listen carefully."

In order to assure that no one has lost his place, the voice on the record tells what the next item is after each 15 items. For example, after item 75, the voice says, "The next item is No. 76."

At the conclusion of part I, 1 minute is allowed for a cleaning-up of answer sheets—that is, erasing or adding any marks that were not filled in properly during the test. Such a cleaning-up period is necessary when answers must be recorded very quickly. The answer sheets for part I are then collected and those for part II distributed. Following are the directions and sample items for part II:

In part II of the test you will listen to a series of words, and mark your answer sheet according to two new rules. Lists of simple three- and five-letter words will be read to you from a phonograph record. There will be 15 words in a list. You must decide whether the first letter of a word comes earlier or later in the alphabet than than the first letter of the word before it. Only words beginning with the letters A, B, C, D, or E will be used.

Three-letter words of one syllable will be considered short words.

Five-letter words of two syllables will be considered long words.

Notice on your answer sheet that after each item there are five spaces, numbered 1, 2, 3, 4, and 5. Your task will be to listen to each word, and then to blacken-in space 1 or 2 according to the following rules which you will be given plenty of time to learn:

Blacken space 1—

A. When a short word begins with a letter earlier in the alphabet than the first letter of the word before it—or;

B. When a long word begins with a letter later in the alphabet than the first letter of the word before it.

Blacken space 2 in all other cases. Be sure to blacken space 2 for the first word in each list of 15 words.

Look at the sample list of words below.

Sample List:

*d*itch —Blacken 2. Blacken space 2 for the first word in a list.
*b*at —Blacken 1. Short word. B of bat comes earlier in the alphabet than d of the preceding word ditch.
*d*izzy—Blacken 1. Long word. D of dizzy comes later in the alphabet than b of bat.
*e*ar —Blacken 2. Short word. Rules A and B do not apply.
*a*pple—Blacken 2. Long word. Rules A and B do not apply.
*c*arry—Blacken 1. Long word. C comes later than a of apple.
*e*nd —
*b*ag —
*c*atty —
*a*llow—

During the test you will hear the list of words introduced and read to you as follows: Begin with item number 1. (Pause one second after each word): Ditch—Bat—Dizzy—Ear—Apple—and so on to the end of the list of 15 items. Then you will be told to start with item 16, and a new list will begin.

(3) *Scoring.*—The scoring formula is R−W+100.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 21.1.

TABLE 21.1.—*Distribution constants for Test of Attention, CI659AX1 based upon a sample of 610 Pilots*[1]

Score	M	SD
Part I	140.3	27.6
Part II	143.6	27.5
Total	284.0	48.0

[1] In class 44II. Tested at Psychological Research Unit No. 3 in February, 1944.

(2) *Reliability coefficient.*—A sample of 507 pilot students yielded the estimates of reliability given in table 21.2.

TABLE 21.2.—*Estimated reliability coefficients by the alternate-forms[1] method for Test of Attention, CI659AX1 based upon a sample of 507 pilots[2]*

Score	r'_{tt}	r_{tt}
Part I	0.79	0.88
Part II	.80	.89

[1] Each part was split into two equal halves.
[2] In class 44H. Tested at Psychological Research Unit No. 3 in February, 1944.

(3) *Part-score intercorrelation.*—For the same sample of 507 pilots the part I-part II correlation was 0.49.

(4) *Difficulty.*—Based upon item analysis of the responses of 532 pilots in class 44H, the test yielded a mean proportion of correct responses of 0.71, corrected for chance, with a range from 0.49 to 0.95 and a standard deviation of 0.09.

(5) *Intercorrelations.*—This test has never been factor-analyzed. Its correlations with some of the tests in the classification battery are shown in table 21.3.

TABLE 21.3.—*Correlations of Test of Attention, CI659AX1 with some classification tests based on 282 classified pilots[1]*

Test	r
Dial and Table Reading, CP621-622A	0.45
Discrimination Reaction Time, CP611D	.43
Mathematics A, CI702E	.34
Reading Comprehension, CI614H	.32
Instrument Comprehension, CI616B	.31

[1] In classes 44G and 44H. Tested in February 1944 at Psychological Research Unit No. 3.

(6) *Test validity.*—Validation results based on a sample of pilots are given in table 21.4.

TABLE 21.4.—*Validity data for Test of Attention, CI659AX1 based upon graduation-elimination of 610 pilots from elementary training[1]*
(p_e=0.89)

Score	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]
Part I	140.7	137.1	27.6	0.07	0.12
Part II	143.6	143.3	27.8	.01	.06
Total score	284.3	286.4	48.0	.04	.10

[1] In class 44H. Tested at Psychological Research Unit No. 3, February 1944.
[2] Assuming an unrestricted stanine standard deviation of 1.87.

(7) *Item validity.*—Validation of individual items of this test disclosed the results recorded in table 21.5.

TABLE 21.5.—*Validity of items of Test of Attention, CI659AX1, based upon graduation-elimination of 532 pilots from elementary training[1]*
(p_e=0.87)

Score	$M\phi$	$SD\phi$	Range of ϕ	
			Low	High
Part I	0.025	0.052	−0.14	0.14
Part II	.015	.064	−.14	.26

[1] In classes 44H and 44I. Tested at Psychological Research Unit No. 3.

(8) *Effect of seating.*—Since the items are orally presented, an analysis of the effect of placement of seats was completed. No significant differences in scores were found among examinees in different parts of the room.

Evaluation.—A satisfactorily reliable test of attention was modeled after two of Wittenborn's tests of attention. Intercorrelations with tests in the classification battery suggest verbal, numerical and space content, although the last is difficult to rationalize. The test probably has no unique pilot validity and has very little validity attributable to known factors.

Following Directions Test, CP402A [8]

This test was constructed early in the program in the Office of the Air Surgeon, Headquarters, Army Air Forces, because it was felt that the ability to carry out orders and follow directions is important to the successful completion of air-crew training.

Description.—The entire test is printed on a standard IBM answer sheet. It consists of alternate lines of text and answer spaces. There are also two panels of dials similar to those found in the cockpit of an airplane. The test contains instructions for filling in the answer spaces. These instructions are constantly altered and supplemented. Figure 21.2 illustrates a section of the test.

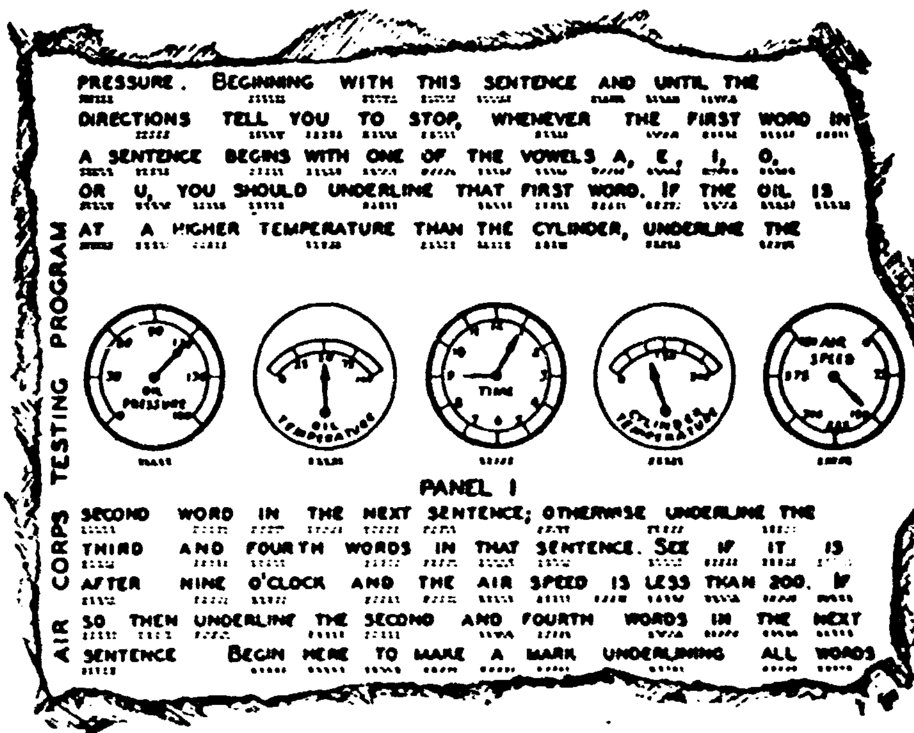


FIGURE 21.2
SECTION OF FOLLOWING DIRECTIONS,
CP402A

[8] Developed in the Office of the Air Surgeon, Headquarters, Army Air Forces. Chief contributors: Lt. Col. Paul M. Fitts and Staff.

(1) *Administration.*—The introductory directions are printed on the IBM answer sheet upside down from the rest of the test so that the examinee will not get a preview of the test while the directions are being read. Ten minutes are allowed for the 420 items.

(2) *Scoring.*—The scoring formula is R−W.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics are given in table 21.6.

TABLE 21.6.—*Distribution constants for Following Directions, CP402A*

Group	N	M	SD
Unclassified aviation students[1]	246	20.7	12.4
Pilots[2]	266	29.7	13.2

[1] Tested in March 1942 at Psychological Research Unit No. 1.
[2] Tested in April and May 1942 at Psychological Research Unit No. 3.

(2) *Reliability coefficient.*—For a sample of 308 unclassified aviation students, tested in June 1942 at Psychological Research Unit No. 1, the test-retest correlation (24 hours interval) was 0.75.

TABLE 21.7.—*Validity data for Following Directions, CP402A, for pilots and navigators (graduation-elimination criterion)*

Group	N_t	p_g	M_g	M_e	SD_t	r_{bis}
Pilots in primary training[1]	547	0.60	5.88	5.73	1.98	0.05
Do[2]	2,658	...	...	...	...	.13
Pilots in basic training[2]	1,942	...	...	...	...	.01
Pilots in primary training[3]	2,505	...	...	...	...	.11
Navigators[4]	228	.79	35.9	30.2	12.7	.26
Do[5]	183	.84	35.2	27.7	16.8	.24
Do[6]	367	.90	21.6	15.8	6.9	.43

[1] Using scaled scores with a mean of 6.00 and a standard deviation of 2.00. Tested at Psychological Research Unit No. 3 in April 1942.
[2] In class 43D. Tested at Psychological Research Unit No. 3.
[3] Class not reported.
[4] New aviation cadets in classes 42-10 through 42-17, Southeast Training Center. Tested at Psychological Research Unit No. 1 in February, March, and April 1943.
[5] Reclassified pilots. Classes and testing data as in footnote [4].
[6] New aviation cadets and reclassified pilots tested at Psychological Research Unit No. 3 from April 1 to August 14, 1942. Class not reported.

(3) *Factorial composition.*—The most significant loadings were found in the integration II (following-directions) (0.54), numerical (0.25), verbal (0.26), and visualization (0.17) factors. The communality was 0.54. For a fuller picture of the factorial composition of this test, see appendix B.

(4) *Test validity.*—Validation results based on several samples are given in table 21.7.

Evaluation.—The Following Directions Test, CP402A, defines a factor called integration II in the analysis of the integration battery (see ch. 10). The test's small validity for pilots in primary training (mean of 0.11) is partly unique, since it appears that this factor has a small loading in the pilot criterion. The test has moderate to high navigator validity, which cannot be fully accounted for by the factors other than integration II. The test is worthy of further development unless a more promising test of the factor can be found. It appears to be some kind of a set or attention factor and may possibly be identified with Wittenborn's attention factor.

TESTS OF ATTENTION UNDER DISTRACTION

Attention Test, CP408A [a]

This test is designed to yield scores for a number of traits: Ability to resist distraction, ability to observe other data while performing a central task, concrete vs. abstract imagery, air-crew interest, and confidence. It is based on the assumptions that the need to perform two tasks in a time-span too brief for either will produce distraction, and that attention is directed toward certain objects rather than toward others as a result of interest. It assumes that pilots must be able to divide attention, be concrete-minded, and interested in airplanes and related objects, and that navigators must be able to resist distractions, be abstract-minded, be interested in scientific objects, and be confident in the results of their own arithmetic when working under pressure.

Description. (1) *Internal characteristics.*—The test consists of three units. Each contains a series of arithmetic problems, comparable to those in the Numerical Operations Test, CI701A, intermingled with distraction words and pictures selected on the basis of job analysis as typical objects involved in air-crew activities and interests. The answers given to arithmetic problems are checked as right or wrong by the examinee. This is followed by a recognition test for the distraction material. The examinee indicates on the following 5-point scale whether or not the item mentioned appeared as distraction material in the numerical operations section of the test:

If you think the item appeared and you are confident of the fact, mark under A.

If you think the item appeared but you are not confident of the fact, mark under B.

If you think the item did not appear and are confident of the fact, mark under C.

If you think the item did not appear and are not confident of the fact, mark under D.

Use E only in those few cases where you simply cannot mark one of the other answers.

(2) Figure 21.3 illustrates a section of the test.

(3) *Administration.*—The test is given in two sections. Section I consists of numerical-operations problems, with distraction material (pictures and words) spaced throughout on the same page. Examinees are informed that their memory for the distracting materials will be tested. Section II contains a memory test on the distraction material.

(4) *Scoring.*—The following are the formulas for the various scores: for confidence, $(R+W)/(R-W)$; for memory, $R-W$; for the interest score for bombardiers, pilots, and navigators, $R-W$; for the numeri-

[a] Developed at Psychological Research Unit No. 1. Chief contributors: Lt. John K. Hemphill, Capt. Donald E. Super.

703120--47--36

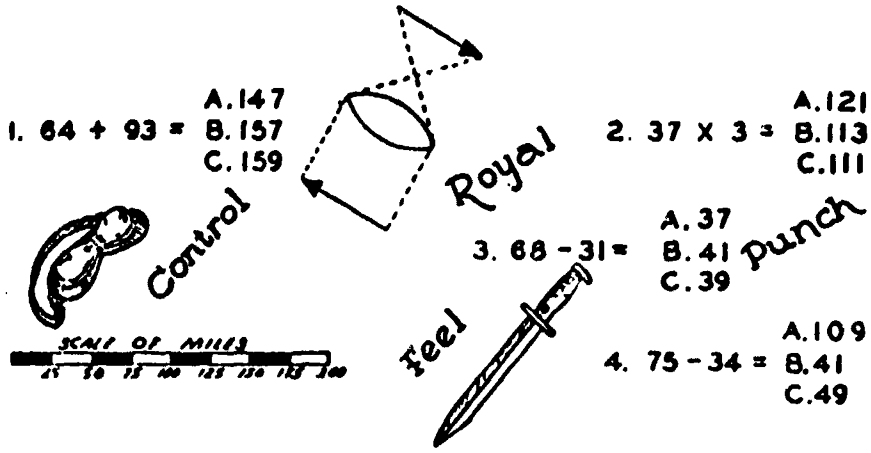


FIGURE 21.3
ITEMS OF ATTENTION TEST, CP 408 A, AND WORK-IN-FLIGHT, CE415A

cal operations score, R−W/2; for Concrete vs. Abstract Imagery, (R−W)/(R′−W′), where R and W refer to correct and incorrect responses for pictures (concrete material), and R′ and W′ refer to correct and incorrect responses for letters (abstract material).

Statistical results.—Results are presented for scores on confidence, memory, pilot interest, navigator interest, numerical operations, abstract vs. concrete memory. The number of cases in the bombardier sample was too small to warrant analysis.

(1) *Reliability.*—Reliabilities of the various scores are shown in table 21.8.

TABLE 21.8.—*Estimated reliability coefficients (odd-even method) for the various scores of the Attention Test, CP408A, based upon a sample of 465 unclassified aviation students*[1]

Score	r_{tt}
Memory for distraction material	0.91
Arithmetic under distraction	.87
Confidence in recall	.97
Concrete memory	.89
Abstract memory	.91
Navigator interest	.79
Pilot interest	.79

[1] Tested in June 1942 at Psychological Research Unit No. 1.

(2) *Intercorrelations of various scores.*—Intercorrelations of some of the scores of the Attention Test, CP408A, are shown in table 21.9.

TABLE 21.9.—*Intercorrelations of certain scores of the Attention Test, CP408A*[1]

Scores correlated	r
Arithmetic vs. memory	0.13
Memory vs. confidence	.34
Concrete memory vs. abstract memory	.46
Pilot interest vs. navigator interest	.48

[1] N not reported. Probably same sample as in table 21.8.

(3) *Validity.*—Tables 21.10 and 21.11 present validity data for the various scores of the Attention Test, CP408A, for pilot and navigator samples.

TABLE 21.10.—*Validity data for the various scores of the Attention Test, CP408A, for a sample of 297 Pilots, based upon graduation-elimination from elementary training*[1]
[(p_e=0.77)]

Score	M_p	M_e	SD_t	r_{bis}
Confidence	58.70	56.33	17.02	0.08
Memory	40.08	39.83	16.82	.01
Pilot interest	14.59	15.04	7.09	−.04
Numerical operations	33.11	34.76	9.75	−.10
Abstract-concrete memory	23.13	21.18	14.63	.08

[1] Tested at Psychological Research Unit No. 1. Class not reported.

TABLE 21.11.—*Validity data for the various scores of the Attention Test, CP408A, for a sample of 61 navigators, based upon graduation-elimination from navigator training*[1]
[(p_e=0.85)]

Score	M_p	M_e	SD_t	r_{bis}[2]
Confidence	56.42	47.56	17.35	0.28
Memory	41.71	38.67	19.44	.09
Navigator interest	13.15	14.06	7.03	−.07
Numerical operations	37.18	34.00	8.82	.20
Abstract-concrete memory	23.35	18.67	13.85	.18

[1] Tested at Psychological Research Unit No. 1. Class not reported.
[2] For these data, the standard error of a zero biserial r is .20.

Evaluation.—The most promising score is the confidence score, for navigators. While the group is very small, the biserial correlation of 0.28 indicates a possibility of a valid measure. In addition, in view of the unusualness of the confidence score, such validity would probably prove to be unique. The nature of the underlying variance is still unknown. The first step would be to verify the navigator validity in a much larger sample. The unusually low navigator validity for the numerical operations score indicates that the sample may not be typical.

Work-In-Flight Test, CE415A[8]

The Work-In-Flight Test is a shorter and simpler modification of the Attention Test, CP408A. It was designed as a test of ability to do arithmetic problems under conditions of distraction and pressure. This ability was thought to be important to the navigator, both as a student in test flights and as an officer in combat and other missions. Additional difficulty and pressure are added to the arithmetic task by (1) scattering irrelevant distracting drawings around the page, (2) testing the examinee for his recall of the irrelevant material and informing him that he is to be so tested, and (3) verbal threat in which the examinee is told of the importance of the test to air-crew work and of the poor performance of many of those present.

[8] Developed at Psychological Research Unit No. 1. Chief contributors: Lt. John K. Hemphill, Capt. Donald E. Super.

It should be pointed out that the measurement of an aviation student's ability to stand up under pressure is rendered difficult by his status. The very fact that he is taking classification tests is a cause for worry on the part of the aviation student, as has been demonstrated by interviews. The desire for a particular type of assignment (generally pilot) and the fear of elimination are widespread. Contrasting a test given with pressure and without pressure in this situation is therefore virtually impossible; tests without pressure are, at best, only tests with pressure slightly reduced.

Description.—There are two parts to the test, each consisting of arithmetic problems with distraction material, followed by a recognition test for the distraction material. Directions emphasize the need for speed, accuracy, and memory. A section of the test is illustrated in figure 21.3.

Effect of pressure.—The Work-In-Flight Test, CE415A, was administered to three samples of aviation students to study the effect of pressure. It was given to 263 aviation students in the standard maximum pressure manner, to 112 students as an arithmetic test only, with the same time limits as in the standard manner, and to 114 students as a memory test only, with the standard time limits. This made it possible to compare (1) mean and variability of arithmetic and memory scores under conditions of minimum and maximum pressure and (2) correlation of the arithmetic score on this test with scores on the simple Numerical Operations Test, CI701A, for minimized and maximized pressure. Minimum pressure consists of doing only one of the two tasks of the test at a time, under the surroundings of aviation-cadet classification testing. The pictures on the test blank are, of course, still present and may serve as a source of distraction. Maximum pressure involves the additional pressure of two simultaneous tasks and of verbal threats.

Results.—The effect of added pressure upon the means and standard deviations is shown in table 21.12.

TABLE 21.12.—*The effect of pressure on the means and the standard deviations of scores of the Work-In-Flight Test, CE415A, administered to 263 aviation students in the standard manner, 112 aviation students as an arithmetic test, and 114 aviation aviation students as a memory test*[1]

Score	Mean		SD	
	Minimum pressure	Maximum pressure	Minimum pressure	Maximum pressure
Arithmetic	18.8	13.5	5.8	5.0
Memory	52.8	24.8	17.2	13.4

[1] Tested at Psychological Research Unit No. 1, about May, 1942.

The correlation coefficients between Numerical Operations, CI701A, and the arithmetic scores of the Work-In-Flight Test, CE415A, with maximum pressure and minimum pressure, respectively, are 0.73 and 0.70, using the samples described above. The reliability (test-retest) of the numerical operations test is 0.81. Kuder-Richardson reliabilities of 0.73 for the arithmetic score with minimum pressure and 0.66 for the arithmetic score with maximum pressure were found, again using the samples described above. The correlations between Numerical Operations and the two arithmetic scores of Work-in-Flight, corrected for attenuation in both variables, are 0.95 and 0.97. The functions measured by the three tests, therefore, are almost identical.

Evaluation.—It is clear that pressure, as here defined, results in a definite lowering of scores on both arithmetic and memory sections of the test and in a lessening of the spread of scores. This is especially true of the memory section. The arithmetic-computation test is, for all practical purposes, a numerical test, with either maximum or minimum pressure. It appears that the distraction material, while influencing scores, does not introduce new variance into a computation test.

Vocabulary Pressure Test, CE201A [a]

This test attempts to measure division of attention under pressure. The task is that of a conventional vocabulary test. Distraction is provided by having the student count the number of auditory signals transmitted at irregular intervals through individual head phones.

Description. (1) *Internal characteristics.*—The test uses the 7 equated vocabulary scales of the ACE Cooperative Vocabulary Test, form R (4), of 30 items each.

(2) *Administration.*—The test was administered in a code-instruction room. While doing the vocabulary test, the students were required to count the number of signals transmitted to them through individual head phones. The seven scales were administered as follows: (1) unspeeded, 6 minutes; (2) moderately speeded, 4 minutes; (3) moderately speeded, 4 minutes; (4) speeded, 4 minutes, with a distraction task added (39 irregularly spaced signals were transmitted on a prearranged time schedule controlled visually by a stop watch); (5) the same as part 4, with 28 irregular signals; (6) excessively speeded, 2 minutes, with instructions to hurry faster given in an urgent voice every 15 seconds; (7) the same as part 6.

(3) *Scoring.*—Separate scores were obtained for the following conditions: (1) normal rate, (2) moderately speeded rate, (3) speed and distraction, (4) excessive speed, and (5) distraction (error score).

Statistical results.—Critical ratios were computed for differences among the means of scores. None were significant. In addition, the following statistics are available.

(1) *Distribution statistics.* Typical examples of distribution statistics are given in table 21.13.

[a] Developed at Psychological Research Unit No. 1. Chief contributors: Capt. Frederick B. Davis and Lt. Col. Laurance F. Shaffer.

TABLE 21.13.—*Distribution constants for Vocabulary Pressure Test, CE201A, based upon a sample of 353 pilots*[1]

Score	M	SD
Normal rate	20.6	8.4
Moderately speeded	35.5	15.1
Speed and distraction	32.5	14.8
Excessive speed	24.2	11.8
Distraction (error score)	4.6	4.2

[1] In class 42E, Southeast Training Center. Tested at Psychological Research Unit No. 1.

(2) *Test validity.*—Validation results based on a sample of pilots are given in table 21.14.

TABLE 21.14.—*Validity data for Vocabulary Pressure Test, CE201A, based upon graduation-elimination of 359 pilots in primary training*[1]
(p_g=0.63)

Score	M_g	M_e	SD_t	r_{bis}
Normal rate	21.08	19.77	8.35	0.10
Speed	36.08	34.48	15.09	.07
Speed and distraction	32.23	32.01	14.78	.03
Excessive speed	24.82	23.17	11.81	.09
Distraction (error score)	4.66	4.60	4.22	.01

[1] In class 42E, Southeast Training Center. Tested at Psychological Research Unit No. 1.

Evaluation.—This test was unsuccessful in its attempt to set up a stress situation. Performance under stress conditions did not differ significantly from performance under normal conditions. None of the scores had a significant pilot validity.

Maze Coordination Test, CM118A [7]

This test attempts to obtain, with a printed test, results comparable to those obtained from psychomotor tests of two-hand coordination.

Description. (1) *Internal characteristics.*—The test consists of three sheets of 8½-inch by 10-inch paper, on each of which are mimeographed two simple maze patterns. After a practice period with the first sheet using each hand separately, the scored trials are made using both hands simultaneously, tracing one maze with the right hand and the other with the left hand. Sixty seconds are allowed for each page. Figure 21.4 shows a page of this test.

(2) *Scoring.*—The score is the number of maze openings passed through, allowing a maximum score of 100 per maze.

Statistical results. (1) *Distribution statistics.*—A sample of 159 pilots yielded a mean score of 49.2 and a standard deviation of 9.5. These examinees were tested at Psychological Research Unit No. 1 in July 1942.

(2) *Reliability coefficient.*—By the alternate-forms method (second maze vs. third maze), an estimated reliability coefficient of 0.90, corrected for length, was obtained. This figure is based on a sample of 579 pilots, tested in February 1943 at Psychological Research Unit No. 1.

[7] Developed at Psychological Research Unit No. 1.

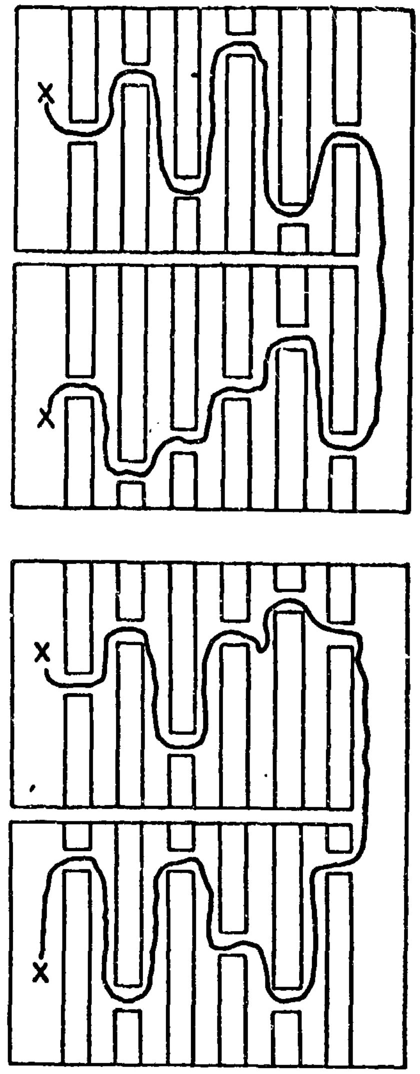


FIGURE 21.4
SAMPLE ITEM OF MAZE COORDINATION, CM118A

(3) *Test validity.*—Validation results based on two samples are given in table 21.15.

TABLE 21.15.—*Validity data for Maze Coordination, CM118A, based upon graduation-elimination from primary pilot training*

N_t	p_s	M_s	M_t	SD_t	r_{bis}
[1]156	0.66	50.91	46.45	9.33	0.29
[2]666	.71	84.3	80.8	16.2	.13

[1] Tested at Psychological Research Unit No. 1. Class not reported.
[2] In classes 42H and 42I. Tested at Psychological Research Unit No. 1.

Evaluation.—Intercorrelations are not available to test the hypothesis that this test is comparable to a psychomotor test of two-hand coordination. It is, nevertheless, a reliable test with moderate validity for predicting graduation-elimination from primary pilot training.

CHANGE OF SET TESTS

A battery of change-of-set tests was constructed by the Medical and Psychological Examining Unit No. 9, Buckley Field, Colo. It was hypothesized that a trait of flexibility of attention contributes significantly to a trainee's success or failure in flying training.

Each test consists of three parts: Part I serves to establish a set in the examinee; the problems in part II are solvable either by the method developed in part I or by a simpler method; the problems in part III are solvable either by the methods used in parts I and II or by a still simpler method.

Internal-consistency statistics are not presented for these tests, because, in general, they would be meaningless, since it is the logical relationship of a given item to its adjacent items and to the whole test which justified its inclusion within the test (no item per se purports to measure the same thing which the test as a whole measures). Test-retest reliabilities are also suspect. If the test were split into equivalent forms, or another form of the test were administered under similar conditions, the effect of learning (insight into the test) would bias any estimate of reliability.

The quantitative scores derived from these tests arise out of the individual differences in the number of problems necessary to break the original set and to approach subsequent items without the predetermined set. Once the examinee has learned to discard his original set, it is expected that he will solve all subsequent problems in the most efficient manner.

Arithmetic Problem Solving Test, CI216A [8]

Description.—In this test the examinee is first presented with seven arithmetic problems that can be solved only by the use of a comparatively complicated procedure. These problems are designed to establish a set. Next he is presented with a series of five similar problems that can be solved either (1) by the comparatively complicated procedure or (2) by a comparatively less complicated procedure. Next, a third series of five similar problems is presented that can be solved either (1) by the comparatively complicated procedure, (2) by the comparatively less complicated procedure, or (3) by a very simple procedure. Of the last ten problems, five may be solved by one of the two more complicated procedures, and five by all three procedures. Although all the procedures are correct, mathematically, the simpler procedures are more economical and efficient. Thus, the primary purpose is not to test for arithmetical ability, but to determine the procedures utilized in solving problems fol-

[8] Developed at Medical and Psychological Examining Unit No. 9. Chief contributor: Staff/Sgt. Milton Rokeach.

lowing the initial establishment of a mental set by means of similar problems.

(1) *Administration.* The following is quoted from the directions to the test:

This is a test of your ability to work arithmetic problems quickly and accurately. All of your problems on this test are based upon the following situation. You are to imagine that you are an army doctor in Alaska. You have just received a shipment of several thousand cubic centimeters (cc.'s) of pneumonia serum in a sealed vessel, and you do not expect another shipment for some time. This sealed vessel has a small faucet so that the serum may be released through the faucet, but no serum can be poured back into the vessel. Your problem is to administer, with the smallest amount of waste, various quantities of serum to soldiers stricken with pneumonia * * *

Look at problem 2:

	Size of containers	*Serum Needed*
2.	17 cc. 37 cc. 6 cc.	8 cc.
	A. 31 cc.	
	B. 29 cc.	
	C. 37 cc.	

* * * you have three empty containers and the sealed vessel. The capacity of these three containers is 17, 37, and 6 cc.'s respectively, the problem being to get exactly 8 cc.'s of serum from the sealed serum vessel * * * with a minimum of waste * * * The correct solution is to fill the 37 cc. from the sealed vessel, pour off 17 cc.'s from the 37 cc. container by filling the empty 17 cc. container, leaving 20 cc.'s in the 37 cc. container. Then pour off 6 cc.'s into the 6 cc. container leaving 14 cc.'s in the 37 cc. container. Again, pour off another 6 cc.'s into the 6 cc. container from the 37 cc. container leaving 8 cc.'s, which is the desired amount, in the 37 cc. container. The simplest way to indicate your solution is thus: $37-17-6-6=8$. Since you have drawn off 37 cc.'s from the sealed vessel and have used only 8 cc.'s, 29 cc.'s have been wasted * * * Your answer should have been B.

The reader will note that the instructions attempt to create a set to solve the problems by the formula b-a-2c, where a, b, and c refer to the first, second, and third containers. The first five test-items can be solved only in this way. The alternative procedures, which can be used in later problems are a–c and c alone.

The time limit for the test is 18 minutes.

(2) *Scoring.*—Since it would be impossible to machine-score this test if the examinee were simply asked to indicate his method of solution, it was necessary to introduce the idea of wastage into the technique. Since the amount of serum wasted depends upon the method of solution, it was possible to construct a multiple-choice item adaptable to machine-scoring. The score is the number of right responses, i. e., choosing the simplest and most economical procedure.

Statistical results. (1) *Distribution statistics.*—A sample of 561 unclassified aviation students (tested at Unit No. 9) yielded a mean score of 16.2, a standard deviation of 7.8, and a range from 0 to 25. The distribution was approximately U-shaped. About 18 percent of the examinees obtained scores of five and below, indicating that they never changed their set. About 26 percent had scores of 24 or 25, indicating that they did shift their set promptly.

There is a marked increase in the incidence of correct answers from one item to the next, indicating that more examinees shift their set and solve the problems by the more efficient method as they progress within part II and part III.

Arithmetic Speed Test I and II (no code number)[9]

Description.—Test I consists of seven parts of simple arithmetic problems involving addition, subtraction, and multiplication. Part I consists of practice problems in addition, subtraction, multiplication, and division. Part II consists of addition problems; part III, subtraction problems; part IV, multiplication problems; and parts V, VI, and VII consist of mixed problems of addition, subtraction, and multiplication.

Test II is contained in parts VIII and IX. Part VIII is a series of simple addition and subtraction problems. Part IX has the same problems, but the examinee's task is to subtract wherever there is a plus sign and to add wherever there is a minus sign.

(1) *Administration.*—Twenty-five seconds are allowed for the 45 items in each part of test I. Ninety seconds are allowed for the 90 items in each part of test II.

(2) *Scoring.*—The change-of-set score for test I is the algebraic difference between the number of problems worked correctly in parts II, III, and IV; and in parts V, VI, and VII. The score for test II is the algebraic difference between the number of problems worked correctly in part VIII and part IX.

Maze Tracing Speed Test (no code number)[10]

Description.—This change-of-set test consists of a series of mazes, the examinee being required to trace the shortest and most direct path from the starting point to the finishing point. In the first series of problems designed to establish a set, the only pathway is a left-going and devious pathway. In the second series of problems, there are two solutions; the left-going, devious solution, and a clearly shorter solution. In the third series of problems, there are three solutions; the left-going, devious solution, the clearly shorter solution, and a very short solution. Figure 21.5 shows a maze of the two-path type with solutions indicated.

(1) *Administration.*—Twelve minutes are allowed for the 29 items.

Code Deciphering Test (no code number)[11]

Description.—This change-of-set test consists of three parts. In parts I and II the examinee is taught two different simple codes. The purpose

[9] Developed at Medical and Psychological Examining Unit No. 9. Chief contributors: Sgt. Clarence Shepherd, Pvt. Garth J. Thomas.

[10] Developed at Medical and Psychological Examining Unit No. 9.

[11] Developed at Medical and Psychological Examining Unit No. 9. Chief contributors: Sgt. Clarence Shepherd, Lt. Francis A. Wimora.

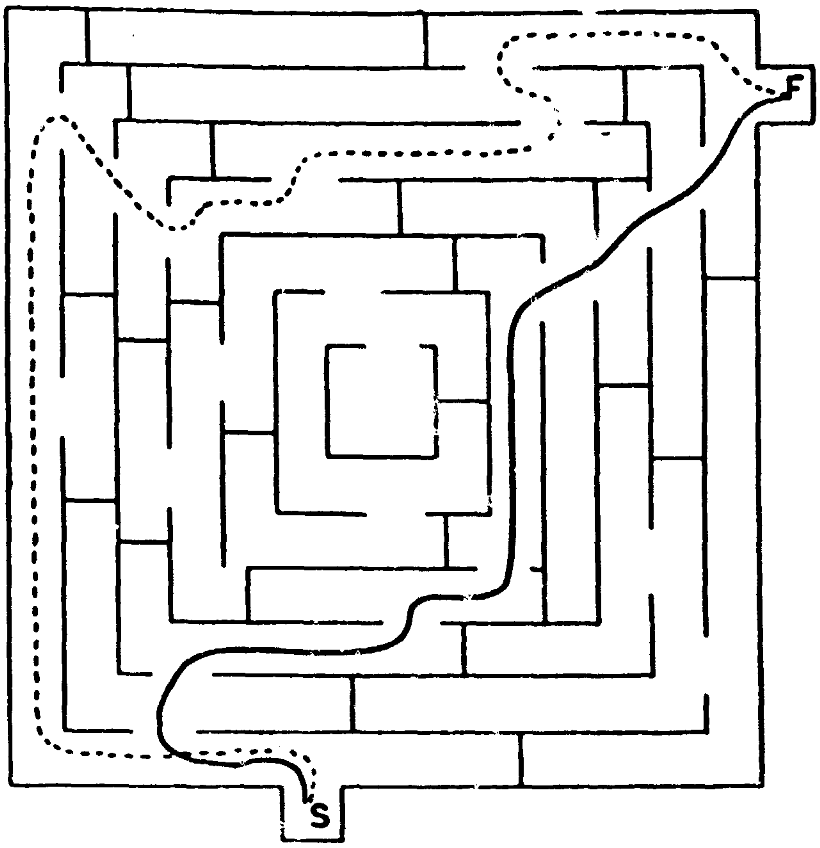


FIGURE 21.5
SAMPLE PROBLEM OF MAZE TRACING SPEED

is to establish sets for two simple codes. In part III, words of both codes appear randomly in the same message.

(1) *Internal characteristics.*—The codes in parts I and II are extremely simple. In part I, the last three letters of the coded word must be placed in front of the first two letters. Thus the coded phrase "alnav ftera llshe ormaj dsroa" is read "naval craft shell major roads." In part II, the last two letters must be placed in front of the first three. Thus, "verco eepst opesl eadah" becomes "cover steep slope ahead." In part III, the words in a message must be solved by both codes. Thus, in the code "kstan inbeg ongal perup gerid" the 1st, 2d, and 5th words are decoded by the method of part I, and the 3rd and 4th words by the method of part II.

(2) *Administration.*—One and one-half minutes are allowed for each part.

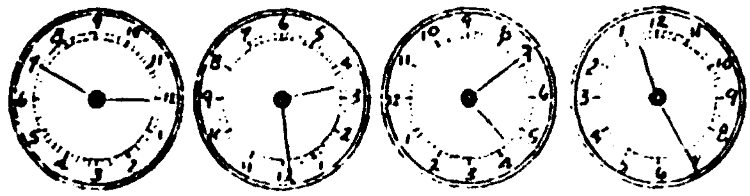
(3) *Scoring.*—The change-of-set score is the difference between the number of words correctly solved in part III versus parts I and II.

Reversed Clock Test (no code number)[10]

Description.—This test consists of 5 parts, each part having 20 items

[10] Developed at Medical and Psychological Examining Unit No. 9. Chief contributors: Tech./Sgt. Robert W. Dietsch, Sgt. Clarence Shepherd.

in the form of replicas of clock faces with different times indicated on each clock face. In parts I and II, the subject reads the time in a normal manner. In part III the hands of the clock are reversed; i. e., the longer hand represents the hours, and the shorter hand represents the minutes. In part IV the figures on the clock faces are placed in counter-clockwise rotation, and the hands move in a counter-clockwise direction, the short hand indicating the hour and the long hand indicating the minutes. In part V the numbers of the clock are in counter-clockwise order, and the hands move in a counter-clockwise direction; also, the longer hand becomes the hour hand, and the shorter hand becomes the minute hand. In each part of the test, the examinee has a specified length of time to finish. Figure 21.6 illustrates an item of this test.



41-A 7:00	42-A 12:17	43-A 9:14	44-A 7:03
B 3:35	B 11:34	B 7:24	B 12:40
C 7:05	C 12:19	C 4:36	C 9:14
D 12:35	D 6:45	D 7:22	D 12:35
E 5:00	E 3:02	E 10:20	E 6:20

FIGURE 21.6
SAMPLE ITEMS OF REVERSED CLOCK, AND CLOCK READING, CP527A

(1) *Administration.*—One minute and 15 seconds are allowed for each part.

(2) *Scoring.*—The change-of-set score is the difference in number of problems answered correctly using part II as a base.

Figure Similarity Test (no code number)[13]

The examinee's task in this change-of-set test is to indicate the maximum number of figures which are alike according to certain principles.

Description.—Each of the first 10 problems has 2 figures alike in shading. This is to establish a set for similarity of shading in the 7 figures which comprise each item. Each of the following 10 problems has 3 figures alike, in that they are of the same shape, and also 2 figures alike, in that they are of the same shading. The following 10 problems have 4 figures alike, in that they are of the same shape and area, 3 figures alike, in that they are of the same shape and area, 3 figures alike, in that they are of the same shape, and 2 figures alike, in that they are of the same shading. Figure 21.7 illustrates a problem of this test. In

[13] Developed at Medical and Psychological Examining Unit No. 9. Chief contributors: Sgt. James G. Madden, Tech./Sgt. Harris W. Roberts, and Pvt. Donald S. Whitten.

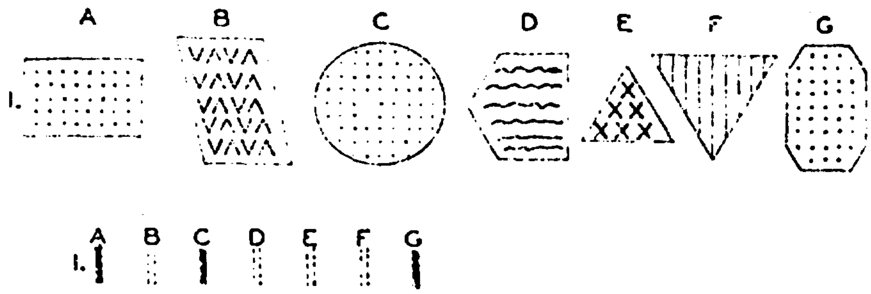


FIGURE 21.7
SAMPLE PROBLEM OF FIGURE SIMILARITY

this sample item, two figures, E and F, are similar in shape; and three figures, A, C, and G, are alike in shading. Since more figures are alike with respect to shading than with respect to shape, the correct answer is the three figures which are alike in shading.

(1) *Administration.*—Ten minutes are allowed for the 30 items.

(2) *Scoring.*—The score is the number right.

Minute Difference Discrimination Test (no code number)[11]

The examinees' task in this change-of-set test is to indicate which one, if any, of four figures is larger than a key figure.

Description.—In the first series of 10 problems designed to establish

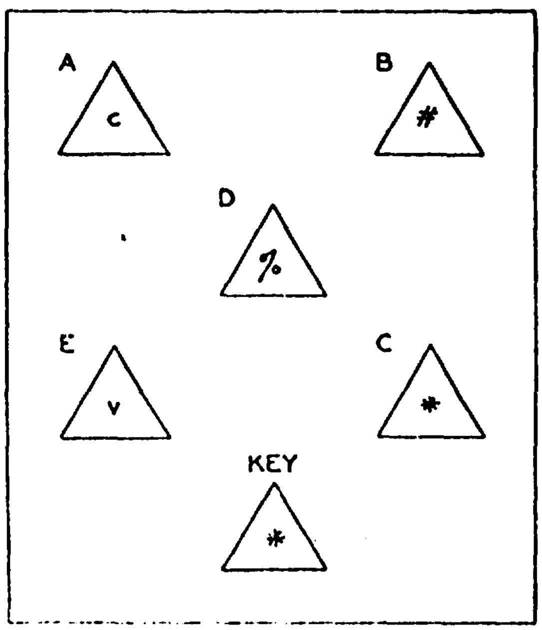


FIGURE 21.8
SAMPLE PROBLEM OF MINUTE DIFFERENCE DISCRIMINATION

[11] Developed at Medical and Psychological Examining Unit No. 9. Chief contributors: Sgt. Robert W. Carlstrand, Pfc. Phillip F. Glucker, and Pvt. Dexter Halpert.

a set, all of the figures in a given position and with a given shading are larger than the key figure. In the next series of eight problems, none of the four figures is larger than the key figure. Then follows another series of four problems in which a set is established for another position. This, also, is followed by a series of eight problems in which none of the figures is larger than the key figure. A series of six items, then, establishes a set for a third position, followed by a series of eight problems in which none of the figures is larger than the key figure. Figure 21.8 shows an item of this test. The score is the number right.

Intercorrelations of change-of-set tests.—The change-of-set battery was administered to 333 unclassified aviation students. One crucial check upon the hypothesis underlying all these tests is found in the intercorrelations among them. If there is a common underlying trait measured by scores on these tests, there should be coefficients of substantial size in the correlation matrix, provided each score is reasonably reliable. The intercorrelations are presented in table 21.16.

TABLE 21.16.—*Intercorrelations of the change-of-set battery based upon a sample of 333 unclassified aviation students*

Test	1	2	3	4	5	6	7	8
1 Arithmetic problem solving	...	0.18	0.10	0.06	−0.02	0.07	0.14	0.13
2 Figure similarity	0.18	...	.07	.06	.08	.16	.04	.14
3 Maze tracing	.10	.07	...	.00	−.01	.17	−.02	.02
4 Arithmetic speed I	.06	.06	.00	...	.04	.02	.04	−.03
5 Arithmetic speed II	−.02	.08	−.01	.04	...	.12	.01	−.05
6 Reversed clock	.07	.16	.17	.02	.12	...	−.03	.06
7 Code deciphering	.14	.04	−.02	.04	.01	−.03	...	.06
8 Minute difference	.13	.14	.02	−.03	−.05	.06	.06	...

It may be seen from the table of intercorrelations that there is no single general change-of-set factor common to these tests. The low intercorrelations may possibly be attributed to low test reliability, but a more plausible hypothesis is that they are due in large part to specificity of function.

Clock Reading Test, CP527A [13]

This test was designed for the purpose of measuring the function or functions involved in responding to a constantly altered or rearranged, but essentially familiar, visual stimulus-pattern. Pilots and bombardiers are often confronted by unexpected changes in stimulus-patterns, which, all-in-all, are familiar to them. It is anticipated, inasmuch as ready adjustment to changes in essentially familiar situations results in an immediate and natural lessening of tension, that the inability to make such adjustments might well be prognostic of susceptibility to combat fatigue and, to a lesser extent, of predisposition to combat neurosis.

Description. (1) *Internal characteristics.*—The test has two parts. Part I consists of 24 diagrams of a conventional clock face. Part II consists of 96 diagrams of clock faces presenting 16 different variations.

[13] Developed at Psychological Research Unit No. 1. Chief contributor: Lt. V. E. Fisher.

The variations are introduced by counter-clockwise arrangement of the hour-numbers, by rotation of the hour-numbers 90°, 180°, or 270°, and by reversing the meaning of the hands. The examinee's task is to tell time under these varying conditions. The test, therefore, is very similar to the reversed clock test, discussed above. Sample items are illustrated in figure 21.6.

(2) *Administration.*—The time limits are 2 minutes and 15 seconds for part I, and 12½ minutes for part II.

(3) *Scoring.*—The scoring formula is R−W/2.

Statistical results. (1) *Distribution statistics.*—Typical examples of the distribution statistics obtained on this test are given in table 21.17.

TABLE 21.17.—*Distribution constants for the Clock Reading Test, CP527A, based upon a sample of 784 pilots*[1]

Score	M	SD
Part I	15.0	4.1
Part II	45.6	17.3

[1] Tested in June 1944 at Psychological Research Unit No. 1.

(2 *Inter-part correlation.*—The correlation of part I with part II gives a product-moment correlation coefficient of 0.39. In view of the difference in content of the two parts, this cannot be considered as an estimate of its reliability.

(3) *Test validity.*—Validation results based on a sample of pilots are given in table 21.18.

TABLE 21.18.—*Validity data for the Clock Reading Test, CP527A, based upon graduation-elimination of a sample of 784 pilots*[1]
$(p_e=0.77)$

Score	M_e	M_s	SD_t	r_{bis}	${}_cr_{bis}$[2]
Part I	15.08	14.52	4.07	0.08	0.13
Part II	46.61	42.03	17.81	.15	.20

[1] Tested in June 1944 at Psychological Research Unit No. 1.
[2] Assuming an unrestricted stanine standard deviation of 2.00.

Evaluation.—Combat criteria are not available to check the hypothesis upon which this test is based. It does, however, have low to moderate validity for the pilot-training criterion. This validity may well be due in part to visualization content, but whether any new valid factor is involved, we do not know. It is most interesting to note the possibility that a test of the simple ability to read clock-faces (part I) may be valid for pilots. This may well be due to perceptual-speed variance in the test.

EVALUATION OF TESTS OF SET AND ATTENTION

The area of attention and set was explored by means of printed tests. The studies, however, were few in number, and no concentrated effort was made in this area. Attention tests were found to have low validity for predicting success in pilot training. On the other hand, they seem to involve intellectual tasks associated with success in navigator training.

The hypothesis that change of set is a fundamental trait that can be measured by means of a battery of tests was not proved to be justified by the results achieved.

The Following Directions Test, CP402A, the only test in the group to be submitted to factor analysis, defined a factor whose identity is not established. Present opinion is inclined to identify it as an integration factor, but this name is offered only as a temporary expedient.

BIBLIOGRAPHY

(1) Cooperative Test Service. *The Cooperative Vocabulary Test: Information Concerning Its Construction, Interpretation, and Use*, New York, 1940.

(2) Gemelli, A. The application of psychological methods to the examination of military aviation candidates, *Arch. Ital. Biol.*, 1917.

(3) Saffiotti, F. U. A brief note on the results of psychometrical research on the problem of the selection of pilots, *G. Med. Milit.*, 1919.

(4) Wittenborn, J. R. Factorial equations for tests of attention, *Psychometrika*, 1943, 8, 19-35.

CHAPTER TWENTY-TWO

Introduction to Temperament Tests[1]

GENERAL CONCEPTIONS

Having accounted for tests in the intellectual and perceptual areas in the two preceding sections of this volume, and having eliminated tests of motor functions from consideration, we have remaining an area of psychological measurement variously known as emotion, temperament, or personality. It is to a general orientation in this field that the present chapter is devoted. Owing to the comparative amorphous character of this field, and to the dangers of confused thinking, a somewhat greater space is given here to a general rationale. There will also be a discussion relating temperament to air-crew requirements and of types of tests designed to help meet the demands for measurement.

Definition of Terms

There will be no attempt to develop and to defend a complete system of psychology of personality. In order that the organization and content of what is to follow may be clearly understood, however, it is necessary to set forth, within the space of a few pages, the outline of a conception of personality. This can be accomplished by means of a few definitions and categorical statements which need not be fully accepted by the reader, but which should serve as a frame of reference for common understanding. The decision to be brief is at the risk of seeming superficial, but the choice is forced upon us. It must be kept in mind in what follows that we are merely concerned with a rational basis for quantitative descriptions of persons. There are other ways of dealing with persons, calling for other types of rationale. Personality is an ambiguous phenomenon, approachable from different points of view.

It should be added that the point of view expressed here is not by any means one adopted by the AAF Aviation Psychology Program. It is believed to be adequate as a framework for presentation of developments arising from diverse theoretical backgrounds. It can be defended on the grounds of convenience for those whose interests are in the development of temperament tests.

Personality.—People are persons because each one is unique. An individual is unique because his pattern of traits or characteristics is different from those of all other individuals. The term "pattern" is used de-

[1] Written by the editor.

703320—47—37

liberately in preference to the alternative of "collection" or any other term implying simple aggregation. *An individual's personality, then, is his unique pattern of traits.*

Traits.—The key to personality, thus conceived, is the general phenomenon of individual differences. When we ask how one individual is different from another, we must get down to particulars. Comparison of one person-as-a-whole with another person-as-a-whole is futile and meaningless. Human comprehension fails to grasp such totalities. Observations and conceptions require abstractions, since human minds are limited in their spans of apprehension. It is not only a convenience but also a necessity for us to compare individuals in one aspect at a time. The process of abstraction, which is required, entails the observation and conception of individuals in a manner that admittedly can be called piecemeal. Analysis is essential whenever observations are made. Two individuals, A and B, are different with respect to aspects c, d, e, f, ... q; individuals B and C differ in aspects e, f, g, ... t; and so on. Any distinguishable way in which one person differs from another is a trait. The term "trait" has thus a very general extension. Trait names are legion and they vastly enrich our language describing people. Traits may be in the nature of common-sense qualities or they may acquire the dignity of scientific sanction and use.

Traits differ in many ways—in scalability; in universality; in generality (v. specificity); in consistency or reliability; in flexibility (v. fixity); in polarity; and in independence (v. dependence).

Some traits are scalable, i. e., each capable of representation by means of a straight line, and some are not. Some traits are either present or absent, e. g., complexes. Others are present to different degrees in different degrees in different persons. The latter are scalable. It is in this type that the psychometrist is primarily interested.

The universality of a trait refers to its extensiveness of manifestation in a population of individuals. Many traits are of such common occurrence, or are held in common by so many individuals, that most people can be ranked on a scale of more or less of those qualities. An example is the degree of total motor activity habitually shown. By means of an appropriate instrument, the total amount of muscular energy expended per pound of body weight during sampled periods of time might be the objective and accurate means of placing individuals on this scale. A trait of somewhat less common extent in a population would be that of marital adjustment (to assume a very abstract variable) which could apply only to those who had opportunity to exhibit behavior describable as good or poor marital adjustment. A still more highly restricted trait would be addiction to ties, since only a small proportion of the population would presumably have ties at all.

The generality of a trait refers to its extensiveness within the individual. Some traits are so general that they pervade almost all the actions of the person and are apparent in almost any kind of action. Such traits are nervousness, meticulousness, and tempo of response. Some examples of more restricted generality are honesty, punctuality, and endurance. These are restricted traits because they appear in special kinds of situations only, and because they represent deviations from special standards of behavior in limited areas. Even more specific traits are platform shyness, liking for beer, and dread of cats or of some particular cat.

By consistency is meant the uniformity with which the same trait (or trait-scale position) is exhibited in repeated (or similar) situations. When an individual is assigned a particular scale position on any trait variable, we can only mean that this is his characteristic scale location (in other words, his mode, median, or mean position). His fluctuation about this point may be great, that of another person may be small. The latter individual is, of course, more predictable on this trait. Some individuals are perennially cheerful, others are constantly depressed, and others run the gamut of the cheerfulness-depression scale from time to time. In some traits the population of individuals can be reliably ranked because few shifts in rank occur from time to time, and in other traits the fluidity of rank positions is such that reliable measurements are almost hopeless and predictions are practically futile. Distinction must be made here between actual shifts of position and changes in manifestation of the trait. Many alterations in an individual's traits are more apparent than real.

The flexibility of a trait refers to its being subject to modifications by learning—in other words, its docility or trainability. This might be regarded as merely one condition for unreliability, but it is a socially important one because it is identifiable and because training is largely controllable. It has no unusual implications for measurement. Its effect is similar to that of any other constant error.

The polarity of a trait refers to whether it is unipolar or bipolar. Most ability traits extend logically from a zero point (complete lack of the ability) to the maximum amount. Most temperament traits are bipolar, each extreme of the scale being given a name of some quality and the two qualities being opposites. A bipolar scale extends through an indifference point or zero point near the middle of the scale.

The independence of traits is a quite important matter in the description of individuals. Traits, as commonly abstracted and named, exhibit various degrees of interrelationship, as noted by direct observation and more clearly by means of intercorrelation procedures. In connection with the goal of trait measurement, it becomes very desirable, from the standpoint of economy and rationality, to discover what the independent or near independent variables of personality are and to measure them separately. In this manner, and only so, can maximum, economical, and meaningful coverage be assured. Having made piecemeal, but accurate, evaluations, we find a knowledge of the interrelationships useful in re-constructing the totality which is the individual. This brings us to an examination of the structure of personality.

The Structure of Personality

In this chapter personality has been defined so broadly as to encompass all individual differences. Personality would therefore include morphological and physiological as well as psychological traits. The morphological and physiological traits are of no concern to us here, though they contribute definitely to making individuals unique in appearance and health.

The psychological traits are of various kinds. One large group—mental abilities—may be subdivided, as they have been in this volume, into intellectual and perceptual categories. The division line is admittedly not completely clear. All other psychological traits may be arbitrarily included under the heading of temperament. Tradition would designate this area alone as personality, using that term in the more restricted sense. It is believed that the break with tradition on this point is in the interests of greater logical clarity. Temperament represents a somewhat amorphous group of traits having at least the one element of emotionality in common. It is becoming recognized that a very important aspect of temperament is that of motivation. Motivation traits include interests and attitudes, areas that have gained attention in vocational and social psychology in recent years, and have yielded to attempts to quantify them. There are probably many who would regard interests and attitudes as coordinate with temperament. The decision on this question is arbitrary and will not be argued here. The desire for a single word other than "personality" to encompass the nonability traits was the deciding factor. A better term than "temperament" is needed.

Within the individual, the structure of personality deserves mention from a somewhat different point of view, namely, the interrelationship of traits. For one thing, intellectual and temperamental traits do not exist separately in the behavior of the individual. Almost any active behavior has both its intellectual and temperamental aspects.

For another thing, some traits seem to be organized in systems or hierarchies. The studies of the areas of reasoning and memory abilities, as related in previous chapters, seem to show that there are widely generalized factors which show themselves in varieties of tests that have fundamental operations in common, and there are also factors of less generality showing themselves in restricted types of tasks within the same general area. A similar situation probably prevails to an even greater extent for temperament traits. Individuals may differ in general proneness to fear (general timidity). Somewhat independently they may differ in general fear of people (general shyness) but not of other objects; in fear of audiences (platform shyness); and in fear of specific persons (for example, conditioned reaction to a dentist). The reason for this is that temperamental traits arise to a large extent through habit formation or conditioning, and they persist according to the laws of conditioning. The laws of generalization and of discrimination determine the generality or specificity of the individual's habits and hence of the qualities resulting from them. The implication from this is that any attempt to predict behavior of a certain quality should take into consideration the generality of the trait in question. Unless the generality of a trait extends to a test situation, it goes without saying that the trait is not susceptible of measurement by tests. Nor are we interested in traits that do not show sufficient universality in the population concerned when selection is to be made from that population.

TEMPERAMENT AND AIR-CREW REQUIREMENTS

In selecting candidates with suitable temperament for air-crew training, several considerations were kept in mind. The selectees must be men who would pass all the hurdles of training, who would successfully carry the fight to the enemy in combat, and who would not break mentally as a consequence of combat. They must be highly motivated not only to become skilled performers in aviation, but also to engage in the duties of soldiers at the front.

Temperament Traits in Training

Job-analysis findings indicate that the chief aspects of temperament that might well be measured are tenseness, nervousness, emotional control, absence of confusion, self-confidence, fear and apprehension, interest or motivation, leadership, and dependability. Some of these are relatively more important in training and others in combat. Without any doubt, emotional traits in general are much more important in combat than in training. This is partly due to the fact that combat officers face more situations that place a premium upon desirable temperament traits. It is probably due to some extent to the fact that the classification tests have never adequately screened out men with weak temperament traits in any way comparable to its screening on aptitudes. The relative dispersions of aptitude and temperament in combat air-crew personnel are therefore quite different, that for temperament being relatively greater and therefore more noticeable.

Reference to chapter 1 will show that, in training, tenseness or nervousness ranked high for bombardiers (sixth place among 16 traits), but only moderately for the navigator (twelfth place among 29), and for the pilot (tenth among 20). Motivation ranked somewhat lower for the bombardier (eleventh among 16) and very low in the list for the navigator (twenty-first among 29), and for pilots in flying training (eighteenth in 20). Self-confidence ranked moderately for both bombardiers and navigators and was not mentioned among the traits for pilots.

Fear of flying seemed to be a very minor cause of elimination from training. For pilots in primary training the data based upon more than

150,000 students showed that only 1.5 percent of all those entering flying training were eliminated for this cause. This represented 6 percent of all eliminees. It is possible that fear played a role in many other eliminations and was not reported as such. For the navigator, the trait of carefulness (or what may even be called meticulousness) seemed to be unique, and it was given very high rank by observers.

Temperament Traits in Combat

In combat, emotional control is rated uniformly high (first place for bombardiers and navigators and second place for pilots among the twenty traits). Motivation is given high rank (third from the top for fighter pilots and seventh place for bomber pilots) for the pilot, but only moderate rank for the bombardier (twelfth place) and for the navigator (thirteenth place). It is probable that many a pilot was enthusiastic about his training due to its civilian potentialities, but was not properly motivated to fly ships in combat.

Dependability ranked second for both bombardiers and navigators, but is in eighth place for pilots. Leadership is rated uniformly slightly above the median for all three specialists. This uniformity is a little surprising in view of the fact that the pilot is most usually placed in positions of leadership. Perhaps the fact that all three are officers and all must at times take command, leads to the expectation of that quality in all alike.

TEST APPROACHES TO TEMPERAMENT

Difficulties in Test Development

In the area of temperament, test development was difficult from several points of view. It is one thing to recognize important traits and to have estimates of their relative importance; it is another thing to devise tests for traits that are not often sufficiently generalized as to appear in test situations. Even if they have that much extension, the control of the test situation is often so difficult that the variable we desire to measure is overwhelmed by other irrelevant variances. It is still another thing to find suitable job criteria to serve for validation purposes. This is even more difficult than in the case of ability tests. As was said before, temperament traits are much more important in combat than in training.

Combat criteria are more difficult to obtain and are unsatisfactory at best. In either training or in combat, the relative variance to be attributed to temperament traits is probably quite small. Emotional failures are spectacular when they do occur, but their occurrence in measureable degree is limited (for example, the 1.5 percent of fear eliminations in primary pilot training). When temperament tests themselves have low variances in particular factors that are valid, extremely low validity coefficients must be expected, and extremely large samples are needed to demonstrate genuine validity.

Another practical difficulty was the very great attrition rate between experimental testing at the time of classification and the completion of a tour of combat duty. Of 10,000 students tested, very few would end up in homogeneous groups with respect to combat criteria. Many would be lost to air-crew training at the time of classification. Consider from this point on only the pilots. Through advanced training many would be eliminated. Of the graduates, some become combat fliers and some not. Of the combat pilots, some fly fighter planes and some bomber planes. In either group some went to the European theater and some to the Pacific. Some completed their missions under one set of conditions and some under another. Validation within any homogeneous combat group is therefore well-nigh impossible if the usual standards are demanded. Add to this the unsatisfactory status of record keeping in combat areas, and the dark picture for combat validation is complete.

Some Principles of Test Development

The types of tests needed to meet the requirements briefly envisaged above had to be selected with several considerations in mind. Because the tests would be used to qualify or disqualify students and to assist in establishing their classification, all possible criticisms as to their fairness, safety, and objectivity had to be forestalled. Controls had to be so adequate that the examinees could not change the fundamental character of the test by failing to undertake the task as the examiner had intended. The right type of motivation and set had to be aroused in examinees. To meet these requirements, a number of principles were generally observed, although explorations over a wide range of possibilities forced violations of those principles at times.

Consideration for the examinee.—The question of safety never arose in connection with printed tests, but was often a factor in performance tests—such as looping chairs, falling hammers, tests of endurance, deprivation of air, etc. It was difficult to frighten or to tire or to otherwise subject the examinee to a severe emotion-inducing situation without risk of injury or of opening the psychologists to the charge of injury even when none occurred. Even "razzing" or insulting or annoying examinees was rarely attempted because of the risk of destroying rapport in other tests and because such practices are generally repugnant in a democratic society. These facts are mentioned merely to show that the program was more or less forced back upon printed tests for its main type of instrument in temperament testing.

In all types of temperament tests, it is usually important to conceal from the examinee the genuine intentions of the tests. In attempting to accomplish this end, great dependence is placed upon the wording of the instructions and upon the disguised appearance of test content. It was one principle of the program to avoid deliberate falsehoods in giving instructions to examinees. This naturally had its restricting influence.

Types of Temperament Tests

Other principles observed can well be pointed out in connection with an account of different types of tests developed or tried out for air crew.

Personality inventories.—From the early days of the program, considerable distrust was evinced concerning personality inventories or questionnaires. This was chiefly on the ground that intelligent examinees can make more favorable scores for themselves than are justified, by falsifying their replies to questions. The extent to which this is true has to be demonstrated for each test and each item. It is possible that some tests of this sort have been so devised, or new ones could be so devised, that the benefits from falsification could be reduced to a minimum. Even if this were true, however, it was felt to be undesirable for any examinee to leave the tests with the conviction that he had been able to outwit the psychologist and to add to his chances of qualification by willfully giving erroneous answers.

The fact that the Bernreuter Personality Inventory had previously failed to predict training success in either CAA studies or in the Army Air Corps also lent a restraining hand. The apparent success of the Shipley Personal Inventory in the armed services as the war progressed, however, and the availability of the experimental testing time were favorable to the administration of a number of commercial personality inventories for validation against pilot training so that extensive knowledge of the true status of usefulness of inventories could be obtained. The studies of these tests were planned in such a manner that if the questionnaire type of item is useful at all in prediction, a large pool of valid items could be selected. Other well-known commercial tests that were included along with the inventories were the Strong Vocational Interest Blank and the Kuder Preference Record.

General information tests.—Other types of printed temperament tests developed in the Army Air Corps were varied in nature. One type which was capitalized upon extensively is the information test. It was seen in chapter 14 how it is possible to assess pilot interest indirectly through what the individual knows (or does *not* know). The same device has been tried in connection with other qualities, such as masculinity and leadership. The information tests have a bona fide appearance to the examinee. The use of information tests for assessing temperamental traits, however, calls for weighting responses empirically, positively or negatively, as they correlate with achievement criteria. Occasionally, examinees have expressed wonderment at the inclusion of certain items. (One might well query why a good pilot should know the correct answer to the question, "An arpeggio is . . . . .") It is desirable to keep such items to a minimum even though they may prove to be predictive. Another qualification to mention regarding general-information tests is that the highly educated individual's score may be biased. One solution to this might be to give a general-vocabulary test in conjunction with the information test, and to use the score as a reference variable or as a suppression variable.

Biographical data.—Biographical data tests were at first somewhat suspect, along with personality inventories. Results which became available during 1942, however, drastically altered the evaluations of this form of test. Item validations yielded scoring keys with quite substantial unique validity, and so a Biographical Data Blank became a permanent part of the classification battery. Early Biographical Data Blanks emphasized factual information regarding previous environment, education, and general experience. It was demonstrated empirically that in spite of the opportunity for a certain amount of falsification under the usual test conditions, the validity of the scores held up. It is possible that with factual information upon which the examinee knows there can be a check-up if necessary, a much higher percentage of truthful responses can be obtained. More recently developed experimental forms of biographical data have resorted to less factual questions, most of which still lack validation studies.

Projective methods.—An attempt was made to get at the deeper recesses of temperament by means of the projective techniques of which the Rorschach and the Thematic Apperception tests are familiar examples. The individual administration of these tests as in clinical practice was recognized as being out of the question for mass testing (some examining units tested as many as 500 men per day with group tests and simultaneously 500 others with psychomotor tests). The Rorschach test was given a thorough and, it is believed, an adequate trial in its approved clinical form, with an attempt to predict success in pilot training from any or all of the data yielded by it. Two variations adapted for group administration (Harrower-Erickson and an improvised form) were also tried out. The Thematic Apperception test of Murray was given in modified form for validation and a number of variations of the thematic approach were developed in printed form for group administration. From none of the projective techniques used were any promising results obtained.

Observational techniques.—Other approaches, hardly to be classified as printed tests but fitting no other category in this series of volumes any better, were observational and interview techniques. Observations were made of examinees during the administration of psychomotor tests and under other situations when they could be observed unobtrusively, and ratings of various traits were recorded by observers. Personal interviews of a substantial length were also held, after which predictions of probable success were recorded. These procedures yielded only the minimum acceptable validities, and all of them suffered seriously from the necessary factors of subjectivity in administration and interpretation.

Motivation tests.—A final group of measuring devices was concerned with interests and attitudes. There is no denying the importance of placing a man in training which he will accept or even welcome. This statement has two facets. On the one hand, does the job provide satisfaction of the man's fundamental interests? On the other hand, is it one that he thinks he prefers? The degree of correlation between genuine interests and self-appreciated interests is unknown. It probably differs for various areas of occupations or even for specific occupations. In the classification of aviation trainees, the approach was from both aspects. It was believed to be important to obtain from each candidate his own statement of preference and to follow it as well as possible in classification, even though a man might not know his own interests well. The use of this datum in classification was weighted far in excess of anything justified by validity against training criteria. Efforts were also made by means of various types of printed tests to assess fundamental interests and morale as factors for success in specialized training. A peculiar instance of the problem of interests was intensively investigated in connection with the assignment of pilots to fighter *v.* bomber training, for in this respect the two assignments apparently had requirements that were significantly different.

The Plan of Presentation of Tests

The organization of the chapters following is open to question in several respects. More than one principle of grouping exists, and few could be followed completely. The result is a compromise of two or more principles.

Chapter 23 begins with personality inventories, most of which are commercial tests. Chapter 24 presents clinical techniques—projective and observational methods and variations thereof. Chapter 25 embraces a miscellaneous list of tests each designed for some particular trait. Chapter 26 is coherent by reason of its single large area of motivation, including interests, morale, and attitudes. Chapter 27 concerns the biographical-data approach. On the whole, reasonably equitable distribution of effort and attention is evident as far as one can justifiably proceed at the present time with printed tests.

The coding of temperament tests.—The grouping of tests in these chapters does not fully agree with the code numbering of the tests. Code numbers were assigned according to the following groups:

100—Absence of tenseness.
200—Absence of nervousness.
300—Absence of fear and apprehension.
400—Temperament.
500—Motivation.
600—Personal information.
700—Projective techniques.
800—Fatigue.

Printed tests are almost entirely confined to the 300 to 700 groups inclusive; since tests of tenseness, confusion, and fatigue are not well adapted to printed form.

Concluding comment.—No strikingly original form of temperament tests was developed by the Army Air Forces, though a number of interesting innovations were tried. Considerable experience was gained in the use of traditional forms and of new variations of the same. Perhaps the most profitable finding was that much can be gained by the adaptation of general-information tests and of biographical-data information in the prediction of success. Such internally heterogeneous tests must be constructed to fit almost every type of job empirically, validating each item against a proficiency criterion. They will probably be found ample, thus developed, in a great many spheres of vocational selection and guidance. They do not satisfy the psychometrist who likes to know what traits he is measuring and who prefers unique tests.

CHAPTER TWENTY-THREE

Personality Inventories[1]

INTRODUCTION

In this chapter, and the next, the results of the validation of many commercially provided instruments for the evaluation of personality will be considered. Inventories and questionnaires are considered in this chapter; clinical-type instruments are considered in the next. The distinction between these two types of instruments is thought to be one of considerable importance. In the clinical-type evaluation, the examinee responds to a relatively unstructured situation, in which he is free to give responses as he sees fit. In the personal inventories, the approach is deliberately controlled as much as possible, in order to force the examinee to react in one specific manner or another so as to obtain the best possible estimate of individual differences on some trait and to minimize differences due to uncontrolled factors.

The primary responsibility for the development of tests of emotion and temperament was assigned to Psychological Research Unit No. 1. In addition to developing many special tests and techniques to assess personality characteristics which were indicated by job analyses to be important to air-crew success, it was decided to undertake the administration and validation of as many commercial tests as time and circumstances permitted. This decision was based primarily upon two considerations: (1) It seemed desirable to estimate the predictive value of various commercial tests, already constructed, both for information as to the actual worth of the instrument in predicting success in air-crew positions, and to obviate any possible postwar criticism that use of comparatively well known instruments was not attempted; and (2) it was hoped that by means of item-analysis procedures (see ch. 3) a large pool of items might be built up from various tests which could be combined into one or more highly predictive tests for use in the selection program.

The criterion utilized almost exclusively for the validation of these personality tests was that of graduation-elimination from primary training. The consensus of flying instructors seemed to be that various attributes of personality are of importance in successfully completing the various stages of flying training. It is obvious, however, that combat criteria are at least equally desirable, since many of the tests purport to assess temperamental variables which should be of value in estimating

[1] Written by Staff/Sgt. Arthur Z. Cerf.

predisposition to com at neuroses. Such criteria, however, were impossible to secure for the purpose.

Several patterns of evaluation have been employed. Two complete procedures have been to (1) validate the scores provided by already existent keys and (2) item-analyze the test and perform a cross-validation, employing random-halves of an entire sample of papers. For many tests, abbreviated techniques have been employed. For some, item-analyses of random halves of the entire sample have been made; cross-validations were then performed only if the analyses indicated a sufficiently large number of valid items. This procedure provides an indirect estimate also of the validity of the keys provided by the test, for, if the items for a given score are not valid, then the score cannot be valid. A second abbreviated procedure has been to validate the scores only.

The greater number of the tests discussed in this chapter are available in commercial form. Three of the tests to be discussed here, however, were developed in the AAF Psychology Program. They are the Restricted Word Association Test CE702B, the Information Blank (S–C) CE410A, and the Teacher Preference Scale CE426A.

The description of the tests will be in two sections; the first will be devoted to personal inventories, the second, to preference inventories.

PERSONAL INVENTORIES

In this section will be included tests that sample various personality areas. Both normal and deviate areas are explored and evaluated, such as introversion-extroversion, economic status, familial reactivity, and hyp ndriacal tendencies. The order of presentation of the tests is purely arbitrary.

In general, the tests were administered from commercial booklets, with responses being recorded on IBM answer sheets, suitable for machine scoring. Each test, though it may be commercial, has been assigned a code number consistent with the AAF series of tests, for convenience. All were given with time limits that permitted all examinees, or nearly all, to respond to all items.

Information Blank S–C, CE410A [1]

This test represents one of the earliest studies on the efficacy of temperament questionnaires as predictive instruments in the AAF Psychology Program. An important objective in preparing and administering this test was to study the feature of falsification by students in response to questionnaire items when the test is taken in a competitive situation. The validation of the scores was a secondary consideration.

Description.—The questionnaire consists of 60 items, of which 25 are presumed measures of the trait of self-sufficiency *v.* sociability (type S), 25 items of the trait self-confidence (type C), and 10 items of truth-

[1] Developed at Psychological Research Unit No. 1. Chief contributor: Lt. Col. Laurance F. Shaffer.

fulness (type T). The 50 items other than truthfulness were adapted from the Bernreuter Personality Inventory (CE133A) with the aid of Flanagan's factor analysis (13). The 10 truthfulness items were written for the questionnaire and are of a type similar to those used in J. N. Washburne's "Social Adjustment Inventory." The truthful response consists of admitting that one doesn't always behave in a particular socially approved fashion.

(1) *Internal characteristics.*—The items are answered either "yes" or "no." Samples of each type of item follow:

Type S: Do you like company when you are feeling sad?
Do you make most of your decisions alone?

Type C: Do you often feel just miserable?
Do you dislike public speaking?

Type T: Did you ever take anything, even a pin, that belonged to someone else?
Did you ever tell what was not quite the truth in order to get yourself out of a difficulty?

(2) *Administration.*—The time required is approximately 15 minutes. Pertinent directions are:

In this blank you are asked some general questions about the way you think. It is not a test. There are no right answers except the answers that tell the truth about yourself.

* * * Work rapidly. Don't think a long time about each question, but record your first judgment promptly. Answer every question. Omit none.

(3) *Scoring.*—The scoring was accomplished on the basis of the author's a priori key. High numerical scores for the three variables indicate self-sufficiency, self-confidence, and truthfulness, respectively.

Statistical results.—Data are available for a sample tested in March 1942 at Psychological Research Unit No. 1.

(1) *Distribution statistics.*—A sample of 200 unclassified aviation students yielded the distribution data presented in table 23.1.

TABLE 23.1.—*Distribution data for the Information Blank (S-C), CE410A, based on a sample of 200 unclassified aviation students*

Scale	M	Median	SD
S	10.7	10.2	2.9
C	18.4	18.6	3.4
T	5.4	5.5	1.8

(2) *Test validity.*—Two sets of validation data are available. Biserial coefficients of correlation between test score and graduation-elimination for the three types of training combined (bombardier, navigator, and pilot) are presented in table 23.2. Bombardiers and navigators were not considered separately because of the small number of eliminees, 3 and 14 respectively. The data for the validity of this instrument in predicting success in pilot primary training alone are presented in table 23.3.

TABLE 23.2.— *Validation data for a group of bombardiers, navigators,*[1] *and pilots,*[2] *using a graduation-elimination criterion for the Information Blank S-C, CE410A*[3]

Scale	M_g	M_e	SD_t	r_{bis}
S	10.83	11.11	2.95	−0.05
C	18.66	18.21	3.52	.07
T	5.35	5.16	1.90	.05

[1] Criterion for bombardiers and navigators is graduation-elimination from advanced training.
[2] Criterion for pilots is graduation-elimination from primary training.
[3] N_t=275, p_g=0.86.

TABLE 23.3.— *Validation data for a group of pilots in primary training, using a graduation-elimination criterion for the Information Blank (S-C), CE410A*[1]

Scale	M_g	M_e	SD_t	r_{bis}
S	10.24	11.29	2.89	−0.21
C	10.03	17.95	3.62	.18
T	5.18	4.67	1.93	.16

[1] N_t=83, p_g=0.75.

For the combined group of bombardiers, navigators, and pilots it would require a biserial coefficient of 0.18 for significance at the 5 percent level and one of 0.24 for significance at the 1 percent level. None of the biserial r's approaches significance at the 5 percent level.

For the group of pilots it would require a biserial coefficient of 0.29 for significance at the 5 percent level and of 0.39 for the 1 percent level. Only the S scale begins to approach significance at the 5 percent level, and this is an inverse relationship, which would seem to indicate that a high degree of self-sufficiency is negatively associated with success in primary pilot training. There was no significant validity found for the truthfulness key.

(3) *Intercorrelations.*—The intercorrelations are r_{sc}=0.10, r_{st}=−0.06, and r_{ct}=−0.02, based on 200 unclassified aviation students. On this basis, it is seen that there is essentially no communality among these variables.

Evaluation.—Validity data obtained for this test indicate that it does not predict graduation or elimination from primary training. None of the correlations differs significantly from zero.

The truthfulness scores of the group were generally low. Of the 10 truthfulness items, the average aviation student made truthful responses to only 5.4 (see table 23.1). This indication of a substantial amount of falsification of response leads to the test author's evaluation of this instrument. "The most striking conclusion of this study is the undependability of the truthfulness of simple questionnaire items, administered in a highly competitive situation of the aviation cadet classification tests. The study throws serious doubt on the possibility of using questionnaires of any kind for the classification of aviation cadets."[8] The near-zero correlations between the truthfulness scores and the other two, however, would leave some doubt concerning the generalizing of the quotation beyond the truthfulness scores themselves.

The Humm-Wadsworth Temperament Scale, CE418A

The Humm-Wadsworth Temperament Scale (10) was developed as a means of evaluating the temperamental qualities of applicants for employment. Accordingly, experimental administration of this instrument was undertaken at Psychological Research Unit No. 1 to evaluate its usefulness in the aviation-cadet selection program.

Description.—The scale is based upon Rossanoff's original theory of personality. "It represents a comparison of the incidence of temperamental traits as they occur in combination within seven groupings or components." (12) The comparison for each component was obtained by contrasting the responses of two groups of subjects to the same test questions. One group included individuals whose behavior manifested the presence of stated traits. The other group included individuals whose behavior was free from manifestations of the component. (10)

A brief description of the various components, as given by the authors of the scale (9), follows:

(*a*) Normal, (N), is primarily a control mechanism, providing rational balance and temperamental equilibrium. It underlies the conservatism and conformity to socially acceptable behavior observed in the well-adjusted person.

(*b*) Hysteroid (H), considered as ethically inferior motivation; domination by considerations of selfish personal advantage; irresponsibility toward the social communty.

(*c*) Cycloid (C), includes emotionality in all dimensions (elevation, intensity, volume); variations in energy, in attention, and in behavior reactions.

(1) Manic (M), is manifested by cheerfulness, activity, alertness, versatility of interest, sometimes irritability.

(2) Depressive (D), components are sadness, inactivity, sluggishness of thought, hopelessness.

(*d*) Schizoid (S), includes imagination, and withdrawal from environment.

(1) Autistic (A), seclusiveness, inward contemplation, narrowed sphere of interests.

(2) Paranoid (P), self-sufficiency, certainty of position, militant defense of ideas, suspiciousness.

(*e*) Epileptoid (E), tendencies toward project-making with inspiration toward achievement. (Some epileptoids have lapses of consciousness or other epileptic symptoms.)

(1) *Internal characteristics.*—The scale consists of 318 questions, each requiring a yes or no answer, of which 159 are scored. Examples of typical questions for several of the components are as follows:

[1] Letter by Lt. Col. L. F. Shaffer, Subject: Exploratory Study of the Application of Personality Questionnaires to Aviation Cadets, dated 22 July 1942.

a. Have your activities ever been interrupted by "blank" or unconscious periods?

b. Are you apt not to give your opinion at a meeting unless you are asked for it, even when you do not like the way things are going?

c. Do you often have to "sleep over" a matter before you decide what to do?

Questions are scored for one or more of the components, different weights being assigned to the various components for different items.

(2) *Administration.*—Between 30 to 90 minutes are required to complete the items.

The directions used are those of the authors:

This set of questions has to do with the way you think. Read each question and answer "yes" or "no." Give the first answer that occurs to you in each case, and let it stand.

(3) *Scoring.*—Raw scores were obtained for each category of temperament. These were sent to Dr. Humm who prepared a profile and a description of the temperamental pattern and who provided a short written summary which contained, by implication, a prediction of success.[4] Definition of several terms seems to be in order.

The profile score for a category is the raw score corrected for no-count and scaled. The no-count is the number of questions answered "no" by the examinee. The difference score is the profile score for a category subtracted from the profile score for the normal.

The weightings for the various raw-score categories are the same as those used in the commercial form of the test.

Statistical results.—Data are available for approximately 200 pilots in classes 43I, 43J, and 43K.

(1) *Reliability coefficients.*—Reliability data are not available for this sample. In the standardization of the scale, the authors determined the reliabilities of the various categories by the split-half method: Normal 0.82; hysteroid 0.85; cycloid manic 0.73; cycloid depressed 0.88; schizoid autistic 0.88; schizoid paranoid 0.70; and epileptoid 0.75.

(2) *Test validity.*—Validity data were computed for raw scores, profile scores, and for profile scores subtracted from the normal components on the profile scores. The results are presented in tables 23.4, 23.5, and

TABLE 23.4.—*Validation data for a group of pilots in primary training, using the graduation-elimination criterion, for the raw scores on the Humm-Wadsworth Temperament Scale, CE418A*[1]

Temperament categories	M_g	M_e	SD_t	r_{bis}
Normal	45.1	44.1	8.82	0.07
Hysteroid	39.7	43.3	12.36	[2]−.18
Manic	25.6	25.8	8.75	−.01
Depressive	25.4	25.0	14.19	.02
Autistic	25.6	22.8	12.76	.13
Paranoid	23.8	23.9	8.90	−.01
Epileptoid	20.8	22.6	5.45	[2]−.21
No-count	177.2	175.6	28.36	.04

[1] $N_t = 202$, $p_g = .63$.

[2] Significant beyond the 5 percent level.

[4] Gratitude is hereby expressed for the generosity of Dr. Humm and for his cooperation in making possible this validation study.

23.6. For these data, a biserial correlation of 0.18 is required for significance at the 5 percent level and of 0.23 at the 1 percent level.

TABLE 23.5.—*Validation data for a group of pilots in primary training, using the graduation-elimination criterion, for the profile scores on the Humm-Wadsworth Temperament Scale, CE418A*[1]

Temperament categories	M_g	M_e	SD_t	r_{bis}
Normal	98.3	97.6	8.04	0.05
Hysteroid	103.0	105.5	8.35	[2]−.19
Manic	95.0	94.9	5.85	.01
Depressive	93.6	93.1	7.21	−.04
Autistic	96.8	94.7	8.15	.16
Paranoid	101.1	101.3	8.93	−.01
Epileptoid	95.6	98.1	7.05	[2]−.22

[1] $N_t=200$, $p_g=0.63$.

[2] Significant beyond the 5 percent level.

TABLE 23.6.—*Validation data for a group of pilots in primary training, using the graduation-elimination criterion, for each profile score subtracted from the normal component on the profile score for the Humm-Wadsworth Temperament Scale, CE418A*[1]

Temperament categories	M_g	M_e	SD_t	r_{bis}
Hysteroid	−3.3	−1.8	5.93	0.16
Manic	−1.8	−1.6	5.78	.02
Depressive	−6.6	−6.5	6.18	.01
Autistic	−7.4	−7.5	5.84	−.01
Paranoid	−2.7	−2.7	8.41	.00
Epileptoid	−7.7	−6.5	5.31	.14

[1] $N_t=200$, $p_g=0.63$.

In treating the profile scores subtracted from the normal component on the profile score, a positive correlation means that having a profile score for a temperament category that is lower than the normal profile score is positively related to success.

(3) *Validity of case summaries.*—It was stated earlier that Dr. Humm supplied written case summaries of the pilot students examined. His conclusions did not include clear-cut predictions of success or failure in pilot training; in validating the summaries, therefore, further interpretation was essential.

Two independent sets of ratings of Dr. Humm's summaries were made. The summaries were sorted into five categories by each of two raters. A rating of five indicates a high probability of success, and a rating of one means a high probability of failure.

The contingency coefficient between these two sets of ratings is 0.77. When converted to make it equivalent to a product-moment correlation, the value is 0.86, which indicates a high degree of agreement between raters.

The distributions of ratings for Dr. Humm's summaries are presented in table 23.7. The five categories, and the number of graduates and eliminees falling into each, are indicated.

TABLE 23.7.—*Distributions of graduates and eliminees according to ratings made on the basis of Dr. Humm's summaries for CE418A*

Rating	Rating I		Rating II	
	Graduated	Eliminated	Graduated	Eliminated
5 (high)	3	1	10	6
4	32	22	35	25
3	35	19	13	6
2	26	11	30	13
1 (low)	24	20	32	23
Total	120	73	120	73

The extent to which the obtained frequencies deviate from the frequencies expected by chance is expressed in terms of chi square. Chi square is 1.05 for rating I and 0.82 for rating II. More than 90 percent of chance deviations would have been as great.

(4) *Validation of integration ratings.*—The cases were also rated according to Dr. Humm's statements concerning temperamental integration. These statements were sorted into three categories, which were described as well-integrated, average integration, and poorly integrated. Two independent ratings were made. The numbers of cases passing and failing for each of these categories of temperamental integration are given in table 23.8.

TABLE 23.8.—*Distribution of graduates and eliminees according to ratings made on the basis of Dr. Humm's statements concerning temperamental integration, for CE418A*

Category	Rating I		Rating II	
	Graduated	Eliminated	Graduated	Eliminated
Well-integrated	21	15	18	14
Average integration	46	28	37	19
Poorly integrated	54	31	66	41
Total	121	74	121	74

For the 195 cases on which data are available, the contingency coefficient between the two ratings is 0.76. When converted to an equivalent product-moment correlation, the value is 0.83.

Again the extent to which the obtained frequencies deviate from the frequencies expected by chance is expressed in terms of chi square. Chi square is 0.11 for rating I and 0.34 for rating II. More than 60 percent of chance deviations would have been as great.

Evaluation.—No significant relationships were found between pilot success in primary flying school and ratings either of Dr. Humm's analyses of temperamental integration or of his case summaries. Only two of the category scores yielded biserial coefficients significant at or beyond the 5 percent level, the hysteroid score yielding a correlation of −0.19, and the epileptoid score of −0.22. If these validities prove to hold up in very large samples, the two scores would no doubt add a small amount to composite prediction, since their contribution would be unique.

Several objections arise in the use of this test. One specific objection to this instrument is its length. Another is the fact that of the 318 items included in the scale, only 159 are scored, which makes it extremely uneconomical. The author's contention, however, is that the unscored items are necessary to insure the validity of those that are scored. The multiweighted scoring is also contrary to efficient test procedures.

Another limitation is phrased in the words of the authors: "As high as 25 percent or 30 percent of normal subjects may invalidate their tests." (10) One important limiting factor in the use of personality tests is the restriction of responses to two categories. It is the stated opinion of both examinees and administrators that lack of a third category "?"—which provides an acceptable answer to an otherwise nonapplicable item—tends to influence adversely the examinee's motivation toward the remaining items in the test. The examinee may easily feel restricted and forced, and if inclined at all to be influenced by pride, will more readily falsify.

The Personal Audit, CE431A

As a commercial test, this instrument was known as the Adams-Lepley Personal Audit. Experimental administration to obtain validation data was undertaken at Psychological Research Unit No. 1.

Description.—This test samples areas of personality that are mentioned frequently in connection with the Rorschach, the Thematic Apperception Test, and other projective techniques. It was felt that this instrument would furnish a more highly structured approach to these various areas of personality than is afforded by projective techniques, and thus serve as a control instrument, the validity of which would be compared with the validities of projective techniques.

(1) *Internal characteristics.*—The test consists of 9 sections of 50 items each. For administrative convenience, the test is divided into three parts. Each of the sections is designed to sample a specific personality area, so that nine areas are tapped.

In part I, scores are derived for sociability, suggestibility, and irritability. In the section on sociability (extroversion), the examinee indicates the degree of his liking for each of 50 activities. Responses are indicated in four degrees: A—like it a great deal, B—some liking for it, C—little liking for it, and D—practically no liking for it. Sample items are: "Raising money for a charity"; and, "Watching a big fire."

In the section on suggestibility (a tendency to agree with authority), the examinee is told that at least half of a group of experts agreed that each of the 50 items is true. Three degrees of responses are afforded: A—complete agreement with decision of experts, B—agree, but with reservations, and C—disagreement with experts. Sample items are: "Majority rule is safest in the long run"; and, "No cultured person would ever use profanity."

In the section on irritability (susceptibility to annoyance), a number of common annoyances is listed. The examinee indicates one of four degrees of annoyance: A—much annoyance, B—some annoyance, C—a little annoyance, and D—never any annoyance. Sample items are: "To have someone read over your shoulder"; and, "To have to wait in a long line to see a show."

Part II contains sections on tendency to rationalize, anxiety, and sexual conflicts.

For each item in the section on tendency to rationalize (tendency to make alibis and excuses) the examinee indicates: A—statement is true, B—doubt the truth of the statement, and C—statement is usually false. Sample items are: "Many good athletes are poor students"; and "Still waters run deepest."

A list of common fears thought to be experienced by all persons to some extent is presented in the section on anxiety. The examinee indicates his degree of fear: A—considerable fear, B—some fear, C—a little fear, and D—no fear. Sample items are: "Having a physician give you the wrong medicine"; and, "Losing your mind or becoming insane."

The section on sexual conflicts consists of a type of controlled word-association test. For each item, the examinee has four alternative responses from which to choose. Sample items follow:

Love A. adore B. esteem C. worship D. yearn
Rape A. attack B. assault C. ruin D. temptation

Part III of the tests contains sections on personal intolerance, flexibility of attitudes, and obsessive thoughts.

In the section on personal intolerance, the examinee indicates the extent of his dislike for certain activities or things: A—a great deal of dislike, B—some dislike, C—little dislike, and D—no dislike. Sample items are: "People who are stingy with their money"; and, "A dirty hobo who asks you for a dime."

The section on flexibility of attitudes contains a list of activities and things for which the examinee is to indicate his present feelings compared with those of 3 or 4 years ago: A—indicates feeling the same as 3 or 4 years ago, B—indicates feeling partly changed, C—feeling is considerably different. Sample items are: "Socialization of medicine"; and, "Capital punishment."

In the section on obsessive thoughts or worry about unsolved problems the examinee indicates the amount of thinking he has done about certain topics. Sample items are: "Kissing or petting between young men and women"; and, "Being demoted or discharged from a job."

(2) *Administration.* Administration is in three parts, to correspond to the parts of the test. Five minutes are allowed for each section, and each new section is begun simultaneously by the whole group. Approximately 45 minutes are required to complete the test. Specific directions

for answering items are given at the beginning of each section. Pertinent general directions are:

This is not a test of your intelligence or of your ability; there are no right or wrong answers. Cadets who have engaged in different kinds of work have a wide range of attitudes and interests. For example, one person might be interested in certain things that may not interest another person at all. In answering the questions in this survey, be careful to avoid indicating one response when you mean another. There are no catch items. Work rapidly, omit no responses.

(3) *Scoring.*—Scoring was accomplished for each of the nine sections by means of the authors' a priori keys. In addition, one of the authors made predictions of pilot success on a nine-point scale.[8] The predictions were based on the test scores, expressed in stanine form. Each stanine score was given a rating of either 1 or 0, so that the nine trait scores could be summated into a total prediction score ranging from zero to eight. For example, for the tests of sociability and flexibility of attitudes, a stanine of four or above became one, and other scores became zero. For the other sections of the test, a stanine of four or below became one in the aggregate weighting.

Statistical results.—The test was validated on a group of 271 pilots in class 43C, originally tested at Psychological Research Unit No. 1 from May 31 to June 27, 1943.

(1) *Test validity.*—As explained above, the clinical predictions on a nine-point scale of success or failure in primary pilot training were validated. These data, along with validity coefficients for the nine section scores, are presented in table 23.9.

TABLE 23.9.—*Validation data for scores and clinical predictions for the Personal Audit, CE431A, for a group of pilots[1] in primary training, using the graduation-elimination criterion*

Score	M_g	M_e	SD_t	r_{bis}
Over-all prediction	5.29	5.54	1.69	−0.08
Sociability	3.91	4.36	1.95	−.12
Suggestibility	4.04	4.07	2.00	−.01
Rationalization	4.12	4.04	1.93	.02
Anxiety	4.05	3.82	1.95	.06
Intolerance	4.11	4.14	1.79	−.01
Attitude flexibility	4.11	4.25	1.88	−.04
Irritability	4.09	3.75	1.90	.09
Sex conflicts	3.94	4.21	1.98	−.07
Obsessive thoughts	3.92	3.89	4.88	.01

[1] $N_t = 271$, $p_e = 0.89$.

The biserial coefficients range from −0.12 to 0.09, which are well within the range to be expected of a change distribution of biserial correlations, the true mean of which is zero. It would require a biserial coefficient of 0.20 for significance at the 5 percent level and a biserial of 0.26 for significance at the 1 percent level.

(2) *Intercorrelations.*—Several intercorrelations were determined between the test author's over-all predictions and the clinical predictions

[8] Predictions made by Maj. William Lepley. The nine-point rating scale is described in chapter 24.

of success made in several of the clinical-procedures techniques (see ch. 24). These intercorrelations are presented in table 23.10.

TABLE 23.10.—*Correlations of author's overall ratings based on the Personal Audit, CE431A, with clinical predictions from other techniques, for a group of pilots in primary training*

Technique	N	r
Interview (CE707A)	189	0.01
Rorschach (CE701-I)	189	.13
Observational Stress Test (CE710A)	188	−.09
Interaction Test (CE425A)	185	−.01
Observation during Psychomotor Test Rest Period (CE709A)	175	−.05

These intercorrelations indicate practically no communality between techniques.

Evaluation.—Because of the low validities −0.12 to 0.09, which could have occurred entirely by chance, for both the over-all ratings and for the trait scores, it is concluded that this test is of no value in predicting pilot performance in primary training.

It is probable that this instrument, along with the majority of personality inventories, does not succeed in predicting air-crew success, because they were not designed specifically for that task. That other verbal temperament tests, not very different in form, but based upon some definite hypothesis concerning air-crew qualities, do succeed, is evidence in support of this conclusion (e. g., see the discussion of the Satisfaction Test, CE409D, in ch. 25).

The Bernreuter Personality Inventory, CE433A

As part of the program of relatively systematic experimental usage of existing commercial personality inventories, this test was also administered at Psychological Research Unit No. 1.

Description.—The standard commercial form of the instrument (2) was used in group-test administration. It consists of 125 items, which are answered "yes," "no," or "?." Typical items are:

Do your interests change rapidly?
Do you usually try to avoid dictatorial or "bossy" people?
Do people ever come to you for advice?

(1) *Administration.*—Twenty-five mintues were allowed for completion of the items. The administrative directions consisted of an explanation of the method of employing the answer sheet and the following general remarks:

The questions on this blank are intended to indicate your interests and attitudes. It is not an intelligence test, nor are there any right or wrong answers.

(2) *Scoring.*—The keys provided for the test were not employed. Instead, an item-validation study was undertaken. It was intended that if a sufficient number of items proved valid, the author's keys would then be validated.

Statistical results.—The data which follow were computed for a sample of 800 pilots tested in June 1944 at Psychological Research Unit No. 1.

(1) *Item validation.*—A distribution of phis is presented in table 23.11, based on data from 600 graduates and 200 eliminees from primary pilot training.[6]

TABLE 23.11.—*Distribution of phis based on validation of responses against the graduation-elimination criterion in primary training for the Bernreuter Personality Inventory, CE433A*[1]

Phi	f
0.08 up[2]	1
0.03 – 0.07	24
–0.02 – 0.02	63
–0.07 – 0.03	16
– ... – 0.08[3]	4
Total	108

[1] $N_t = 800$. $P_s = 0.75$.
[2] The highest positive phi was 0.10.
[3] The lowest negative phi was —0.10.

In interpreting these phi coefficients, it can be said that for an N of 800, a phi of 0.07 is significant at approximately the 5 percent level, and a phi of 0.09 is significant at the 1 percent level of confidence. Altogether, six phis, two positive and four negative, reached or exceeded the 5 percent level of significance. Two attained the 1 percent level of significance.

Evaluation.—On the basis of the very small number of phis attaining statistical significance, it was decided that this instrument contains an insufficient number of valid items for the prediction of primary pilot success to make further scoring measures or statistical analyses worth while.

An Inventory of Factors S T D C R, CE434A

This inventory was developed as a commercial instrument on the basis of factor-analysis studies of various personality-questionnaire items. Experimental administration was undertaken in an attempt to assess the validity of the inventory for pilot selection.

Description.—The factors S, T, D, C, and R taken together probably cover the area of personality generally encompassed by the concept of introversion-extroversion. According to the author (4), each factor actually represents a dimension of personality with two opposite poles. The factors may be described as follows:

S—Social introversion, as exhibited in shyness and tendencies to withdraw from social contacts.

T—Thinking introversion, an inclination to meditative thinking, philosophizing, and analyzing one's self and others.

[6] Only "yes" responses were utilized in tallying the distribution.

D—Depression, including feelings of unworthiness and guilt.

C—Cycloid tendencies, as shown in strong emotional reactions, fluctuations in mood, and tendency toward flightiness or instability.

R—Rhathymia, a happy-go-lucky or carefree disposition, liveliness and impulsiveness.

(1) *Internal characteristics.*—The inventory consists of 175 items, each of which is to be answered by "yes," "?," or "no." Some of the items are scored for only one factor, others are scored for several, as in the sample items which follow:

Sample 1: Do you express yourself more easily in speech than in writing? (If answered "no," this item has an S value of 1.)

Sample 2: Are you inclined to act on the spur of the moment without thinking things over? (If answered "yes" this item has an R value of 2.)

Sample 3: Are you inclined to be moody? (If answered "yes," this item has a value of 1 for factors T, D, and C.)

(2) *Administration.*—Pertinent directions are:

Read each question in the test booklet in turn, think what your behavior has usually been, and mark that answer space, after the corresponding item number, that describes your behavior best. Mark the "?" only when you are unable to decide between the "yes" and "no." Be sure to answer every question. There is no implication of right or wrong in any of these questions * * *

(3) *Scoring.*—The inventory is scored by means of one key for each of the five factors. The keys for factors S, D, and C give all significant responses a weight of one point, and those for factors T and R weight some responses two points.

Statistical results.—Experimental administration of this instrument was completed at Psychological Research Unit No. 1 in May 1944 with approximately 1,100 pilots who took primary training.

(1) *Test reliability.*—Reliabilities were not computed for aviation students. The test author cites reliabilities obtained by combining alternate sixths of the items into two pools of approximately equal lists and the use of the Spearman-Brown formula. This procedure yielded estimated reliabilities of 0.92, 0.89, 0.91, 0.91, and 0.89 for factors S, T, D, C, and R, respectively, in a sample of 200 college undergraduates selected at random from a criterion group.

(2) *Intercorrelations.*—The intercorrelations of the five factor scores are presented in table 23.12.

TABLE 23.12.—*Intercorrelations of the five factor scores obtained on the Inventory of factors S T D C R, CE434A*[1]

Factor	S	T	D	C	R
S	...	0.24	0.58	0.38	−0.51
T	0.24	...	.51	.48	−.21
D	.58	.51	...	.91	.11
C	.38	.48	.91	...	.16
R	−.51	−.21	−.11	.16	...

[1] N_t=1,106 pilots in primary training.

(3) *Test validity.*—The five factor keys published with the test were used for scoring. Validity data are presented in table 23.13.

TABLE 23.13.—*Validation data for a group of pilots in primary training, using the graduation-elimination criterion, for the Inventory of Factors S T D C R, CE434*[1]

Factor	M_p	M_e	SD_t	r_{bis}	$_c r_{bis}$[2]
S	12.63	12.97	7.37	−0.03	−0.06
T	32.54	33.83	8.71	[3]−.09	−.04
D	12.93	13.93	8.17	[4]−.07	−.12
C	19.28	20.38	9.62	[4]−.07	−.12
R	47.23	46.79	10.10	.03	−.01

[1] $N_t=1,106$, $p_p=0.77$.
[2] Corrected to an unrestricted augmented stanine standard deviation of 2.10.
[3] Significant at the 1 percent level.
[4] Significant at the 5 percent level.

(4) *Item validity.*—The distribution of phis based on item analysis data used in the cross-validation study is presented in table 23.14. These data are based only on the "yes" responses.

TABLE 23.14.—*Distribution of phis based on samples of pilots in primary training, using the graduation-elimination criterion, for the Inventory of Factors S T D C R, CE434A*

Phi	*f* (odds)[1]	*f* (evens)[2]
0.13 to 0.17	1	1
0.08 to 0.12	9	9
0.03 to 0.07	27	37
−0.02 to 0.02	50	58
−0.07 to −0.03	43	36
−0.12 to −0.08	19	10
−0.17 to −0.13	1	2
−0.22 to −0.18	1	0
Total	151	153

[1] N=526.
[2] N=516.

For an N of 520, a phi of 0.09 is significant at approximately the 5 percent level, and a phi of 0.12 is significant at the 1 percent level. Sixteen phis in the odds sample (7 positive and 9 negative) reached or exceeded the 5 percent level of significance, with 5 of these (1 positive and 4 negative) reaching or exceeding the 1 percent level. For the evens sample 26 phis (9 positive and 17 negative) reached or exceeded the 5 percent level, and 8 of these (2 positive and 6 negative) reached or exceeded the 1 percent level.

(5) *Cross-validation data.*—Cross-validation data were also computed for this sample, which was split into two groups, odds and evens. Separate item analyses were accomplished for each subsample, and two scoring keys devised. The criteria for scoring a response were: (1) a phi significant at or beyond the 5 percent level (0.09) and (2) a split of 85–15 or better. This is standard procedure for many of the tests to follow in this chapter and in others.

The evens group was scored with the odds key, and the odds group was scored with the evens key. The validities obtained are presented in table 23.15.

TABLE 23.15.—*Cross-validation data based on two groups of pilots in primary training using a graduation-elimination criterion, for the Inventory of Factors S T D C R, CE434A*[1]

Group	Key	Score	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[2]
Odds[3]	Evens ...	Rights[4] ..	14.29	14.38	2.98	−0.02	0.03
		Wrongs ...	5.25	5.20	2.14	.01	−.06
		R−W ...	9.04	9.18	4.60	−.02	.04
Evens[5]	Odds	Rights ...	8.62	8.27	1.99	.10	.10
		Wrongs ..	6.83	6.57	2.55	.06	.11
		R−W ...	1.79	1.70	3.84	.01	−.03

[1] p_g=0.79 for both groups.
[2] Corrected to an unrestricted augmented stanine standard deviation of 2.10.
[3] N_t=526; number of scored items=27.
[4] Rights mean positively keyed responses and wrongs mean negatively keyed responses.
[5] N_t=516; number of scored items=19.

Evaluation.—On the basis of the validity study, it appears that the inventory of factors S T D C R is not promising for predicting graduation-elemination from primary pilot training. The validities of factor scores for D (−0.07) and C (−0.07) are both significant at the 5 percent level and for factor T (−0.09) at the 1 percent level. Thus negative relationships with thinking introversion, depression, and cycloid tendencies probably exist but are too low to be useful in selecting pilots.

There would seem to be an excess number of valid responses beyond the confidence limits, but, in view of the apparent unimodality of the distribution, with its central tendency at zero, and the failure of the cross-validation test (see table 23.15), it would appear that there are very few genuinely valid items for pilot training in the inventory.

The relatively high intercorrelation between several of the factors seems to indicate that there is some overlapping in the items. Thus, for factors D and C, r=0.91, while for D and S, and D and T, the correlations are 0.58 and 0.51. The latter are tolerable and may represent the actual degree of correlation between the factors, but the correlation between C and D is so high as to demand a revision of the keys for these two factors for an aviation-student population. These intercorrelations are quite similar to those obtained by the test author on a college undergraduate sample, with two exceptions: in the college sample r_{DT}=0.15 and r_{CT}=0.14, while the corresponding values for the aviation student population were 0.51 and 0.48, respectively.

The Guilford-Martin Personnel Inventory, CE436A

This commercially provided instrument was designed for two primary purposes. First of all, it was designed to assist supervisors of workers in business and industry in detecting and diagnosing those individuals who are personally maladjusted in their jobs, particularly those who are discontented and likely to become troublemakers. Secondly, the test was designed to extend the list of traits of temperament already assessed by the Inventory of Factors S T D C R, CE434A. The area covered by this inventory may be roughly designated by the term "paranoid," though only the extreme symptoms deserve that appellation borrowed from psychopathology.

Description.—The authors hypothesize that there are several somewhat related aspects of the paranoid disposition. These aspects may be described as (1) subjectivity (taking things personally; ideas of reference; touchiness), (2) belligerence (domineering attitude; craving for superiority), (3) suspiciousness, and (4) faultfinding or hyper-criticalness. In setting up the lists of items diagnostic of these traits, it was found that the last two could not be scored with sufficient independence to justify separate keys. The list of traits measured by the inventory, therefore, reduces to three. Using the names of the more favorable end of the scale in each instance, they are:

O—Objectivity (as opposed to personal reference or a tendency to take things personally).

Ag—Agreeableness (as opposed to belligerence or a dominating disposition and an overreadiness to fight over trifles).

Co—Cooperativeness (as opposed to faultfinding or overcriticalness of people and things).

(1) *Internal characteristics.*—The inventory consists of 150 items. The questions are to be answered by either "yes," "?," or "no." All but eight of the questions yield scores for one or more of the three factors. Examples of questions, one for each of the scoring categories, follow:

O—Are you inclined to be thinking about yourself much of the time? (Answer of "No" is significant for O.)

Ag—Are you annoyed when people tell you how you should do a thing? (Answer of "No" is significant for Ag.)

Co—Does it seem to you that human beings hardly ever learn to avoid making the same mistakes twice? (Answer of "no" is significant for Co.)

(2) *Administration.*—Thirty minutes are sufficient for completion of the 150 items. Essential comments from the directions are:

Read each question in turn, think what your behavior has usually been, then blacken your answer sheet * * * Answer by "?" only when you are unable to decide between the Yes and No. There is no right answer to any of these questions except the answer that tells how you think or feel about it * * *

(3) *Scoring.*—Each item may be scored for one or more of the three factors. In this case, also, the sample was split into odds and evens, or groups. Separate item validations were accomplished on these two subsamples, and two scoring keys were made. The criteria for scoring a response were: (1) a phi significant at or beyond the 5 percent level (0.10) and (2) a split of 90-10 or better. It is interesting to note that there is practically no correlation between the odd and even keys and that only 13 scored responses are held in common.

Statistical results.—Data are available for approximately 950 pilots, originally tested in May 1944 at Psychological Research Unit No. 1, who took primary training.

(1) *Test reliability*.—Reliabilities were not computed for aviation students. Reliabilities (comparable halves) cited by the authors of the test are: O, 0.83; Ag, 0.80; and Co, 0.91.

(2) *Intercorrelations*.—The intercorrelations of the three scores are presented in table 23.16.

TABLE 23.16.—*Intercorrelations of the three scores obtained on the Guilford-Martin Personnel Inventory, CE436A*[1]

Factor	O	Ag	Co
O	...	0.60	0.62
Ag	0.60	...	.62
Co	.62	.62	...

[1] N_t=945 classified pilots in primary training.

The trait-score intercorrelations are high, which is to be expected, as all three traits are attempts to measure paranoid temperament. At the same time they are sufficiently low to provide for differential measurement. These data closely approximate the intercorrelations obtained by the test authors on a civilian sample.

(3) *Test validity*.—Validation results based on a sample of 945 pilots in primary training are presented in table 23.17. The three factor keys published with this test were used.

TABLE 23.17.—*Validation data for a group of pilots in primary training using the graduation-elimination criterion for the Guilford-Martin Personnel Inventory, CE436A*[1]

Factor	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]
O	52.27	50.25	11.78	[3]0.10	016
Ag	30.62	28.70	9.23	[4].12	.13
Co	60.38	56.47	16.19	[4].14	.19

[1] N_t=945, p_g=0.79.
[2] Corrected to an unrestricted augmented stanine standard deviation of 2.10.
[3] Significant beyond the 5 percent level.
[4] Significant at or beyond the 1 percent level.

(4) *Item validity*.—After dividing the sample of answer sheets into random halves, the responses to the items were correlated with the graduation-elemination criterion from primary pilot training. The distributions of the phi coefficients are shown in table 23.18.

TABLE 23.18.—*Distribution of validity phis based on item analysis of the Guilford-Martin Personnel Inventory, CE436A*

Phi	*f* (odds)[1]	*f* (evens)[2]
0.13 to 0.17	3	2
0.08 to 0.12	12	6
0.03 to 0.07	17	30
−0.02 to 0.02	45	49
−0.07 to −0.03	38	36
−0.12 to −0.08	14	16
−0.17 to −0.13	6	1
Total	135	140

[1] N=466.
[2] N=467.

For an N of 468, a phi of 0.09 is significant at approximately the 5 percent level of confidence, and a phi of 0.12 is significant at the 1 percent level. There are 31 phis in the odds sample (15 positive and 16 negative) that reached or exceeded the 5 percent level of significance, with 14 of these (5 positive and 9 negative) reaching or exceeding the 1 percent level. For the evens sample 19 phis (7 positive and 12 negative) reached or exceeded the 5 percent level of significance; 5 of these (2 positive and 3 negative) reached or exceeded the 1 percent level.

(5) *Cross-validation data.*—Cross-validation data were obtained, as shown in table 23.19.

TABLE 23.19.—*Validity data based on two groups of pilots in primary training, using the graduation-elimination criterion, for the Guilford-Martin Personnel Inventory, CE436A*

Groups	Key	Score	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Odds[2] ...	Evens	Rights[3]	13.45	12.88	2.48	[4]0.13	0.12
		Wrongs	4.40	4.81	1.76	−.13	−.14
		R−W+20 .	29.04	28.12	3.60	[4].14	.14
Evens[5] ..	Odds	Rights	36.33	34.70	7.07	[4].13	.21
		Wrongs	7.01	6.74	4.44	.04	−.03
		R−W+20 .	49.35	47.88	8.74	.10	.19

[1] Corrected to an unrestricted augmented stanine standard deviation of 2.10.
[2] Odds sample N_t=466, p_g=0.79; number of scored items=27.
[3] In this table, rights means positively scored responses, wrongs means negatively scored responses.
[4] Significant at the 5 percent level.
[5] Evens sample N_t=467, p_g=0.79; number of scored items=59.

Evaluation.—On the basis of the validity study, it appears that the Guilford-Martin Personnel Inventory has some promise for predicting graduation-elimination from primary pilot training. Thus, scores for factors Ag (r_{bis}=0.12) and Co (r_{bis}=0.14) are significantly related to the criterion at the 1 percent level and for factor O (r_{bis}=0.10) at the 5 percent level. In view of the fact that these scores undoubtedly contribute something different from the classification battery, they would make some addition to its predictive value for pilot selection.

On inspection, it may be seen that each empirical key is valid, yet little or no correlation was shown between keys. This fact may be due to sampling fluctuation. It is recommended that a combined key be used to score future papers of this test for pilot selection.

Inventory of Factors G A M I N, CE435A

In the factor-analysis approach to the problems of temperament, several traits have been identified, and a series of inventories have been constructed which effectively measure some of these traits. The inventory of factors G A M I N (6) adds five more temperament variables to the eight already measured by the two other tests in this series, namely, the Guilford-Martin Personnel Inventory, CE436A, and the Inventory of Factors S T D C R, CE434A. Hence experimental administration of this instrument, for purposes of determining validity for predicting pilot success, was undertaken at Psychological Research Unit No. 1.

Description.—The definitions of the trait names used in this inventory, as in the preceding two, derive from factorial studies and of subsequent item analyses. The five traits measured here include:

G —General pressure for overt activity, or general activity vs. inactivity.

A —Ascendancy in social situations as opposed to submissiveness; leadership qualities.

M—Masculinity of attitudes and interests as opposed to femininity.

I —Lack of inferiority feelings; self-confidence.

N—Lack of nervous tenseness and irritability, or psycho-somatic stability *v.* nervousness and jitteriness.

(1) *Internal characteristics.*—The inventory contains 270 items, each calling for a response of "yes," "?," or "no." Examples of items from the keys for each trait follow:

For trait G.—"Are you inclined to be quick in your actions?"
"Can you turn out a large amount of work in a short time?"

For trait A.—"Do you find it difficult to get rid of a salesman to whom you do not care to listen or give your time?"
"Have you ever, on your own initiative, organized a club or group of any kind?"

For trait M.—"Do you like love scenes in a movie or a play?"
"Do you (or would you) like to go hunting with a rifle for wild game?"

For trait I. —"Do you feel that the average person has made a better adjustment to life than you have?"
"Do you feel confident that you can cope with most situations that you will meet in the future?"

For trait N.—"Do you often become irritated over little annoyances?"
"Do you have nervous habits such as chewing your pencil or biting your fingernails?"

(2) *Administration.*—All examinees finish in approximately 45 minutes. Pertinent directions are:

> • • • Read each question in turn, think what your behavior has usually been. Then on your answer sheet blacken the space that describes your behavior best. Be sure to answer every question. There is no right answer to any of these questions except the answer that tells how you think or feel about it • • •

(3) *Scoring.*—The authors' published keys were used. Scoring weights had been found for each response to every item by using Guilford's *abac* method (5). This procedure yielded final keys consisting of 41 items for trait G, 50 items for trait A, 52 items for trait M, 69 items for trait I, and 69 items for trait N. Only nine items are scored for more than one trait.

Statistical results.—Experimental administration of this instrument was completed in June 1944 at Psychological Research Unit No. 1 with approximately 780 pilots who took primary training.

(1) *Test reliability.*—Reliabilities were not computed for aviation students. The authors' estimated reliabilities secured by correlating comparable halves of the keyed items and correcting for length by the Spearman-Brown formula were: 0.89 for trait G, 0.88 for trait A, 0.85 for trait M, 0.91 for trait I, and 0.89 for trait N, for a sample of college undergraduates.

(2) *Intercorrelations.*—The intercorrelations of the five factors are presented in table 23.20.

TABLE 23.20.—*Intercorrelations of the five scores obtained from the Inventory of factors G A M I N, CE435A*[1]

Factor	G	A	M	I	N
G	....	0.50	0.09	0.26	0.10
A	0.50	...	.32	.60	.44
M	.09	.32	...	.50	.51
I	.26	.60	.50	...	.75
N	.10	.44	.51	.75	...

[1] N_t=782 pilots in primary training.

These intercorrelations are quite comparable to those obtained on the authors' original civilian data. Inspection of table 23.20 shows that factor G is independent of factors M and N. Factors A and I, and I and N show a fairly high degree of intercorrelation, yet the correlations are still low enough for each factor score to provide differential measurement. The remaining intercorrelations are moderately low, which is indicative of some success in the measurement of independent traits.

(3) *Test validity.*—Validation data based on a sample of 782 pilots in primary training are presented in table 23.21.

TABLE 23.21.—*Validation data for a group of pilots in primary training, using the graduation-elimination criterion, for the Inventory of Factors G A M I N, CE435A*[1]

Factor	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]
G	38.18	37.44	10.87	0.04	0.07
A	66.07	67.34	14.53	−.05	−.03
M	56.76	56.67	9.63	.01	.04
I	88.83	89.65	16.78	−.03	.00
N	83.37	84.14	17.72	−.03	−.01

[1] N_t=782, p_p=.76.
[2] Corrected to an unrestricted augmented stanine standard deviation of 2.10.

A biserial correlation of 0.10 is required for significance at the 5 percent level and of 0.13 for significance at the 1 percent level. None of the correlations approaches significance at the 5 percent level.

(4) *Item validity.*—After dividing the sample of answer sheets into random halves, the responses to the items were correlated with the graduation-elimination criterion from primary pilot training. The distribu-

703320—47—39

tions of the phi coefficients are presented in table 23.22. Only "yes" responses were used in tallying these distributions.

TABLE 23.22.—*Distribution of validity phis based on an item analysis of the Inventory of Factors G A M I N, CE435A*

Phi	f (evens)[1]	f (odds)[2]
0.18 to 0.22	0	1
0.13 to 0.17	2	6
0.08 to 0.12	16	20
0.03 to 0.07	61	42
−0.02 to 0.02	83	75
−0.07 to −0.03	34	46
−0.12 to −0.08	11	21
−0.17 to −0.13	3	3
Total	210	214

[1] $N=384$. [2] $N=375$.

In interpreting these phi coefficients, it can be said that for an N of 380, a phi of 0.10 is significant at approximately the 5 percent level and a phi of 0.14 is significant at the 1 percent level of confidence. In the odds sample, 30 phis (17 positive and 13 negative) exceed the 5 percent level of significance, and 6 of these (3 positive and 3 negative) reach or exceed the 1 percent level. For the evens sample, 15 phis (8 positive and 7 negative) reach or exceed the 5 percent level of significance, and 5 of these (2 positive and 3 negative) reach or exceed the 1 percent level.

(5) *Cross-validation data.*—A cross-validation study was made, using two empirical scoring keys. The usual criteria of a split better than 85–15 and significance at or beyond the 5 percent level of confidence were employed in constructing the scoring keys. The resulting data are presented in table 23.23.

TABLE 23.23.—*Validity data based on two groups of pilots in primary training, using the graduation-elimination criterion, for the Inventory of Factors G A M I N, CE435A*

Group	Key	Score	M_p	M_f	SD_t	r_{bis}	$_cr_{bis}$
Odds[1]	Evens	Rights[2]	16.71	17.00	2.93	−0.10	−0.10
		Wrongs	7.31	7.40	3.08	−.07	−.05
		R − W + 20	49.43	49.81	4.92	−.05	−.02
Evens[3]	Odds	Rights	13.25	12.96	3.16	.06	.03
		Wrongs	25.79	25.01	5.51	.08	.09
		R − W + 20	14.13	14.34	7.68	−.02	−.04

[1] Odds sample N_t = 375, p_e = 0.75; number of scored items = 51.
[2] In this table, rights means positively scored responses, wrongs means negatively scored responses.
[3] Evens sample N_t = 384, p_e = 0.75; number of scored items = 66.

For an N of 375 or 384 a correlation of 0.14 is required for significance at the 5 percent level and of 0.18 for significance at the 1 percent level.

Evaluation.—On the basis of the obtained validities, it appears that the Inventory of Factors GAMIN holds practically no promise as an instrument for predicting graduation-elimination from primary pilot training. None of the five categories gives a biserial correlation that even closely approximates significance at the 5 percent level. This is interesting from one point of view, namely, the failure of any sign of validity for the masculinity score. This result should be considered in connection with the hypothesis regarding masculinity mentioned in chapter 25.

No collection of items based on item correlation with the criterion seems any more promising. There would seem to be an excess number of valid items beyond the confidence limits, but, in view of the apparently unimodal distribution with its central tendency at zero and the failure of the cross-validation test (see table 23.23) to show significant biserial correlations, it is probable that there are few, if any, genuinely valid items in this collection for the prediction of primary pilot training success. Probably the item validities merely represent a random sampling around zero.

It might be that this instrument would be of some use if the scores obtained here were combined into a composite profile, together with the scores obtained on the inventory of factors S T D C R and the Guilford-Martin Personnel Inventory. The scores for the 13 factors might be plotted on a composite graph, by means of which significant profile configurations of traits would be revealed. Such a study would prove interesting and possibly fruitful.

The Minnesota Multiphasic Personality Inventory, CE437A

This study was designed to determine whether the scores which can be derived from the group form of the Minnesota Multiphasic Personality Inventory are related to success in flying training and to predisposition to combat fatigue. Administration for validation purposes was undertaken at Psychological Research Unit No. 1 in July 1944.

Description.—The Minnesota Multiphasic Personality Inventory (8) is an instrument that attempts, in one test, to provide scores on all of the more important phases of personality. The 550 items are arbitrarily classified under 26 headings as follows:

1. General health (9 items).
2. General neurologic (19 items).
3. Cranial nerves (11 items).
4. Motility and coordination (6 items).
5. Sensibility (5 items).
6. Vasomotor, trophic, speech, secretory (10 items).
7. Cardio-respiratory system (5 items).
8. Gastro-intestinal system (11 items).
9. Genito-urinary system (5 items).
10. Habits (19 items).
11. Family and marital (26 items).
12. Occupational (18 items).
13. Educational (12 items).
14. Sexual attitudes (16 items).
15. Religious attitudes (19 items).

16. Political attitudes—law and order (46 items).
17. Social attitudes (72 items).
18. Affect, depressive (32 items).
19. Affect, manic (24 items).
20. Obsessive and compulsive states (15 items).
21. Delusions, hallucinations, illusions, ideas of reference (31 items).
22. Phobias (29 items).
23. Sadistic, masochistic trends (7 items).
24. Morale (33 items).
25. Items primarily related to masculinity-femininity (55 items).
26. Items to indicate whether the individual is trying to place himself in an improbably acceptable light (15 items).

(1) *Internal characteristics.*—This group form of the test consists of the same 550 items that were originally developed as an individually administered, card-sorting form. The response to each item is recorded in one of three categories: (a) "true, or mostly true," (b) "not usually true, or entirely untrue," and (c) "cannot say." Sample items are:

I like to read newspaper editorials.
Someone has it in for me.
I do not like everyone I know.

(2) *Administration.*—Pertinent directions are:

In this inventory you are asked for information about your feelings, your likes and dislikes, and a great many other things. This is not a test. There are no "right" answers except the answers that tell the truth about yourself. To a large extent, your success in air-crew training will depend upon how well you are understood by those in charge. It is therefore to your own interest to fill out this blank carefully and completely.

(3) *Scoring.*—The authors have prepared keys for hypochondriasis, hysteria, depression, psychasthenia, psychopathic deviation, masculinity-femininity, paranoia, schizophrenia, and hypomania. It was decided to use these keys only if an item-validity study indicated that these keys had potential validity for prediction of pilot success in primary training.

Statistical results.—Data are complete for 856 pilots in primary training.

(1) *Test reliability.*—Reliability data were not computed. The authors have found that the test-retest reliability coefficients of their keys range from 0.71 to 0.83.

(2) *Item validity.*—The sample of pilots was split into two groups, odds and evens, each having an N of 400, and separate item analyses were accomplished for the 2 subsamples. Distributions of phis for the odds and the evens group on the basis of a split of 90–10 or better is presented in table 23.24. Only positive phis are given, since the test proved to be essentially a set of two-choice items for this population.

TABLE 23.24.—*Distribution of phis based on two samples of 400 pilots in primary training, using the graduation-elimination criterion, for the Minnesota Multiphasic Personality Inventory, CE437A*

Phi	f (odds)	f (evens)
0.15 to 0.19	2	6
0.10 to 0.14	26	38
0.05 to 0.09	118	122
0.00 to 0.04	205	182
Total	351	348

For an N of 400, a phi of 0.10 is significant at approximately the 5 percent level of confidence, and a phi of 0.13 is significant at the 1 percent level. For the evens group, 44 phis reached or exceeded the 5 percent level of significance, of which 19 reached or exceeded the 1 percent level of significance. For the odds group, 28 phis reached or exceeded the 5 percent level of significance, of which 5 reached or exceeded the 1 percent level of significance.

Evaluation.—In view of the apparently unimodal distribution of phis with a central tendency at zero it is probable that there are few, if any, genuinely valid items in this collection for the prediction of primary pilot training success. These analyses are taken to indicate that no validity for the test could result from a cross-validation study, and, accordingly, none was attempted.

It is to be remembered that this instrument was developed for use as a clinical device in the prediction and confirmation of diagnoses of clinical entities. It may be that it would prove valid in predicting criteria such as combat fatigue. It was not possible to carry out this type of validation study.

Minnesota Personality Scale, CE438A

The information cited here is concerned entirely with the form for men of the Minnesota Personality Scale (3), which was also administered at Psychological Research Unit No. 1 in January 1944.

Description.—Five aspects of personality are measured: Morale, social adjustment, family relations, emotionality, and economic conservatism. These traits were reported to have resulted from a factor analysis of several personality tests, and they are defined as follows:

a. High morale scores are indicative of belief in the institutions and future possibilities of society. Low scores usually indicate cynicism or lack of hope in the future.

b. High social adjustment scores tend to be characteristic of the gregarious, socially mature individual in his relations with other people. Low scores are characteristic of the socially inept or under-socialized individual.

c. High family relations scores usually signify friendly and healthy parent-child relations. Low scores suggest conflicts or maladjustments in parent-child relations.

d. High emotionality scores are representative of emotionally stable and self-possessed individuals. Low scores may result from anxiety states or over-reactive tendencies.

e. High economic conservatism scores indicate conservative economic attitudes. Low scores reveal a tendency toward liberal or radical points of view on current economic and industrial problems.

(1) *Internal characteristics.*—The test is divided into 5 sections, and it consists of 218 questions. Each item has five alternative responses. Part I consists of questions 1–40 dealing with morale. Each item is to be answered by one of the alternatives: (SA strongly agree; (A) agree; (U) undecided; (D) disagree; and (SD) strongly disagree. Sample items are: (*a*) "Court decisions are almost always just" and (*b*) "There is really no point in living."

Part II comprises items 45–105, and deals with social adjustment. The alternatives from which the examinee is to choose are: (AA) almost always; (F) frequently; (O) occasionally; (R) rarely; and (AN) almost never. Sample items are (*a*) "Do you have a fairly good time at parties?" and (*b*) "Are you able to recover quickly from social blunders?"

Part III is concerned with family relations, and includes items 106–135. These items are answered with one of five alternatives of the series (AA) almost always, through (AN) almost never. Sample items are: (*a*) "Do you and your parents live in different worlds, so far as ideas are concerned?" and (*b*) "Have you had to keep quiet or leave the house to have peace at home?"

Part IV deals with emotional stability, consisting of items 142–176. These items are answered on the same continuum of responses used in part III. Sample items are: (*a*) "Are your eyes very sensitive to light?" and (*b*) "Do ideas run through your head so that you cannot sleep?"

Part V consists of items 186–218 dealing with economic conservatism, answered with alternatives ranging from (SA) strongly agree, to (SD) strongly disagree. Sample items are: (*a*) "Private doctors should encourage trends towards socialized medicine," and (*b*) "The government should take over all large industries."

For purposes of convenience in machine scoring, there are gaps in the numbering of items between parts I and II, III and IV, and IV and V.

(2) *Administration.*—Forty-five minutes suffice for almost all examinees to complete the test. Pertinent directions are:

The following pages contain a number of statements about which there is no general agreement. People differ in the way they feel about the statements, and there are no right or wrong answers. We are trying to study certain aspects of personality that are important in your adjustment to aircrew training. You can help us by answering each question honestly and thoughtfully. Happiness and satisfying achievement are definitely related to your personal adjustments; therefore any effort to study this aspect of your life is worth your cooperation.

* * * mark the one alternative which best expresses your feeling about the statement. Whenever possible let your own personal experience determine your answer * * *

One practice problem is given before beginning the test.

(3) *Scoring.*—The inventory is machine-scored using the authors' keys, with each response being weighted from one to five.

Statistical results.—Results are available for 338 examinees who took primary pilot training.

(1) *Test reliability.*—Reliabilities were not computed for this group. The authors' corrected odd-even reliability coefficients are as follows: Part I, 0.84; part II, 0.97; part III, 0.95; part IV, 0.94; and part V, 0.92.

(2) *Intercorrelations.*—The intercorrelations of the part scores are given in table 23.25. In general, the degree of intercorrelation is sufficiently low to insure some degree of independence of the scores.

TABLE 23.25.—*Part score intercorrelations for the Minnesota Personality Scale, CE438A, based on the scores of a group of 338 pilots in primary training*

	1	2	3	4	5
1. Morale	...	0.44	0.30	0.35	0.52
2. Social adjustment	0.44	...	.42	.58	.27
3. Family relations	.30	.42	...	.64	.17
4. Emotionality	.35	.58	.64	...	.25
5. Economic conservatism	.52	.27	.17	.25	...

(3) *Test validity.*—Validation data, based on a sample of pilots in primary training, using the graduation-elimination criterion, are given in table 23.26. For this sample a biserial coefficient of 0.15 is required for significance at the 5 percent level and a coefficient of 0.20 for significance at the 1 percent level.

TABLE 23.26.—*Validation data using the graduation-elimination criterion for the five categories of the Minnesota Personality Scale, CE438A, based on a sample of pilots in primary training*[1]

Score	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[2]
Morale	47.91	47.65	9.39	0.02	0.04
Social adjustment	66.10	68.94	18.61	−.09	−.10
Family relations	45.50	45.22	9.03	.02	.04
Emotionality	43.16	43.04	11.69	.01	.00
Economic conservatism	15.52	15.09	6.95	.04	.07

[1] $N_t = 338$, $p_g = 0.79$.

[2] Corrected to an unrestricted stanine standard deviation of 2.00.

Evaluation.—On the basis of the present validation study, it appears that the Minnesota Personality Scale, CE438A, has no value for predicting success in primary pilot training. The highest validity coefficients obtained were 0.07 for economic conservatism and −0.10 for social adjustment. These coefficients do not differ significantly from zero.

Shipley Personal Inventory, Format B, CE601B

The purpose of this test is to detect those individuals who exhibit psychoneurotic or psychotic symptoms. Experimental administration of this instrument in the Air Corps was accomplished by Psychological Research Unit No. 1 and by headquarters of the AAF Training Command.

In this report, the results of three validity studies will be mentioned. The first study was designed for the purpose of validation of this test against the criterion of success in primary pilot training. The second and third studies were conducted for the purpose of validating the test against the criterion of satisfactory-unsatisfactory adaptability ratings for military aeronautics, for a group of aviation students and for a group of WASPs.

Description.—The scored items have been divided by the author into 20 clusters, on a purely a priori basis, and they were designed as convenient groupings for the psychiatrist, to help in obtaining a qualitative picture of the individual. In general use of the test, primary concern is with the single quantitative score, and the clustering feature was designed primarily to appeal to the psychiatrically inclined. Without entering into a discussion of their natures, the clusters are: Psychopath A and B, neurotic A and B, irresponsible, inadequate, social poise, sex, sociability, hypochondriasis, near psychosis, gastro-intestinal, epilepsy-dizzy, family stability, family not closely knit, mood swing, school success, femininity, job-school link, and miscellaneous. These categories were not considered in the AAF's use of the inventory. They are presented merely as a means of describing its content.

(1) *Internal characteristics.*—The inventory consists of 145 items, 60 of which are scored to yield the total number of undesirable responses. Each item affords two choices, from which the examinee is to select the one which seems to apply better to him. The choices are printed in two columns, as in the following sample:

L		*R*
I take life easy.	....	I tend to worry.
I like to listen to the radio.	....	I prefer a bang-up party.
I like to stay put.	....	I've gone on the bum.

(2) *Administration.*—Since testing time is not specified in the directions for the test, the first group of 280 tested was used to standardize the time. It was found that approximately 50 percent of the group completed half of the items in 15 minutes and that 80 percent completed the test in 30 minutes. Accordingly, in later administrations, at the end of 15 minutes all examinees were admonished to work more rapidly. Thirty minutes were allowed for completion of the test.

Pertinent directions are:

In this questionnaire you are to give information which will help others understand you. You are to indicate certain things about your job preferences, interests, etc.

In each question you will always have two answers to choose between * * * the one on the left side of the page, and the one on the right. Choose the answer which fits you best. Even if neither fits you very well, you must choose the one that fits you better than the other * * *

Remember, you must always choose one answer to each question. Be sure not to skip any questions. Work rapidly.

(3) *Scoring.*—The test was scored by the author's key, consisting of 60 choices which are considered undesirable. In a preliminary administration of the inventory, the author found that an undesirable score of 18 seemed to be significant as a line of demarcation between the abnormal and the normal group.

Statistical results.—The results of three validity studies and other pertinent statistics, where available, will be presented.

(1) *Test reliability.*—Test reliabilities were not computed.

(2) *Test validity.*—In the first study, a sample of 1,419 pilots originally tested at Psychological Research Unit No. 1 in October 1943 yielded a biserial correlation of 0.06, uncorrected for restriction of range, between "undesirable" scores in this inventory and the graduation-elimination criterion in primary pilot training. The biserial coefficent corrected to an unrestricted stanine standard deviation of 2.00 is 0.05. The mean score for graduates is 6.73, for eliminees 6.41, and the standard deviation for both combined is 3.34. Of this sample 81 percent were graduates. The obtained coefficient is barely significant at the 5 percent level, but it is in the unexpected direction.

The second study employed the criterion of satisfactory-unsatisfactory adaptability ratings for military aeronautics. A sample of 2,107 aviation students tested at Psychological Research Unit No. 1 yielded a biserial correlation of −0.35 between the number of undesirable responses on the test and the criterion of satisfactory-unsatisfactory adaptability rating for military aeronautics. No distribution data are available for this sample. For a sample of this magnitude a biserial coefficient of 0.10 is significant at the 1 percent level (the split was 0.975–0.025).

A comparison was made of the mean scores on the inventory, for 53 aviation students who received unsatisfactory adaptability rating for military aeronautics scores and for 510 who received satisfactory scores. The 510 were members of a random sample obtained from a total group of 2,054 tested at Psychological Research Unit No. 1. The mean score for the sample was 6.71, with a standard deviation of 3.49; the mean for the total group was 6.77, with a standard deviation of 3.49. The mean score for the group of 53 cases was 9.49, with a standard deviation of 4.06. The critical ratio of the difference between means of the satisfactory and unsatisfactory scores is 4.80, which indicates a very significant difference.

A third study pertains to scores of 194 WASP's (classes 44-W-6 and 44-W-7) on the adaptability rating for military aeronautics, as obtained by a psychiatrist interviewer, and scores on a revised format B. This revision of the test was modified so as to omit items that applied only to males and to include new items applicable to females. The ratings were divided into four categories: (*a*) satisfactory, (*b*) borderline satisfactory, (*c*) borderline unsatisfactory, and (*d*) unsatisfactory. The mean undesirable score on the inventory for the entire WASP group is 8.65.

With the WASP criterion data combined into two categories—two satisfactory subgroups in one and two unsatisfactory subgroups in the other—the biserial coefficient between adaptability rating for military aeronautics and inventory score is −0.36. This value is close to that obtained for aviation students.

(3) *Item validity.*—After dividing the aviation student sample of answer sheets into two random halves, the responses to the items were correlated with the graduation-elimination criterion from primary pilot training. Since this test is a two-choice one, only the A responses were tallied in the phi distributions. The distributions of the phi coefficients are shown in table 23.27.

TABLE 23.27.—*Distribution of phis based on item analysis for cross-validation of the Shipley Personal Inventory, CE601B*

Phi	*f* (evens)[1]	*f* (odds)[1]
0.15 to 0.19	1	0
0.10 to 0.14	9	4
0.05 to 0.09	28	26
0.00 to 0.04	53	55
Total	91	85

[1] *N*=636.

In interpreting these phi coefficients, it can be said that for an N of 636, a phi of 0.08 is significant at approximately the 5 percent level of confidence, a phi of 0.11 is significant at the 1 percent level. In the odds sample, 11 phis reached or exceeded the 5 percent level of significance with 2 of these reaching or exceeding the 1 percent level. For the evens sample, 15 phis reached or exceeded the 5 percent level; 4 of these reached or exceeded the 1 percent level. On inspection, it would seem that the number of phis occurring at significant levels could have been expected by chance.

Evaluation.—On the basis of these data, the Shipley Personal Inventory, format B, is not useful as an instrument for the prediction of graduation-elimination from primary pilot training. The biserial coefficient of 0.06 between the criterion of graduation-elimination from pilot primary training and undesirable responses to the inventory is barely significant at the 5 percent level. This coefficient is in the opposite direction from that expected, however, and could be a chance deviation from zero.

The critical ratio of the difference between the unsatisfactory group and the satisfactory group on the adaptability rating for military aeronautics is 4.80, which indicates high significance. The biser[illegible]cient between undesirable responses to the inventory and the criterion of unsatisfactory adaptability rating for military aeronautics is 0.35, which is far above the level of significance required at the 1 percent level, but which is somewhat questionable because of the one-sided split of these data. A similar correlation was obtained with a small group of WASPs, in which the split was well balanced.

Several objections that are of interest arise in the use of this instrument. It is reported that some of the items appear unsuitable for use in the selection program, because they present a choice between an acceptable and unacceptable response (e. g., "I get embarrassed easily"—"I seldom get embarrassed"), or because both items are socially unacceptable and of the type of "Have you stopped beating your wife?" (e. g., "Our family scraps often came after someone had been drinking"—"Drinking never was the cause of our family scraps").

It is reported that during the administration of the test there was considerable laughter and many comments and questions. The comments of the examinees, in general, centered about the fact that some items were silly, that many seemed to be repeated, and that there was difficulty in choosing between the alternatives, with considerable resentment being shown in regard to being forced to make a choice in situations in which the examinees denied having any previous experience.

Restricted Word Association Test, CE702B

The purpose of this test is to predict emotional stability during and after combat service.[7]

Description.—The idea of this test is based on the assumption that words, or more specifically the meanings of words, become associated in conformity with the relationships that exist between the individual's affective attitudes and tendencies to action and the situations to which such subjective factors relate. If, for instance, a given individual's principal tendency to action with respect to Hitler would be to attack, then the stimulus-word "Hitler" would elicit some such response word as anger, rage, or hate. If a given individual is excessively self-concerned, a number of different stimulus-words should elicit responses referring to himself. If an individual is generally negativistic, he might be expected to respond frequently with a word which is the opposite of the stimulus-word. Finally, if an individual is oversensitized to some given fact, stimulus-words denoting the fact or connoting characteristics of the fact should reveal the examinee's sensitivity.

(1) *Internal characteristics.*—The test consists of 50 items. Each item consists of a stimulus-word followed by five alternative response-words. Thirty of the stimulus words were selected as related to the work of air crew. These are mixed with 20 from the Kent-Rosanoff list. The

[7] Developed at Psychological Research Unit No 1. Chief contributors: Lt. Vivian E. Fisher and Capt. Donald E. Super.

examinee is asked to indicate the response-word in each case with which in his thoughts and/or feelings the stimulus-word is most strongly associated. A sample item follows:

In my thoughts and/or feelings FAILURE is most strongly associated with:
- (*a*) Success.
- (*b*) War.
- (*c*) Myself.
- (*d*) Flying.
- (*e*) Fear.

(2) *Administration.*—The test is group-administered. There is no time limit, sufficient time being allowed for all examinees to complete the test. Pertinent directions are:

This is an information test to see how words are associated with other words in cadets' thoughts and feelings. You will be given a key word and below it will be listed five other words. You are to select the one of these five words which is most strongly connected in your thoughts and feelings with the key word.

This is not a test of how well you understand words. There are no right or wrong answers. Different people have different associations between words, and the intention of this test is to get information about these differences.

The choices for each of the key words on this test have been selected in such a way that there is no single best choice or answer for any key word. You will do best if you do not linger over any question, but select the word which you feel, at your first impression, seems to go most strongly with the key word.

(3) *Scoring.*—Test scoring is accomplished by means of an a priori key.

Evaluation.—No statistical data are available to permit an evaluation of this test.

PREFERENCE INVENTORIES

In this section are considered those tests that evaluate interests and preferences. The general line of approach is somewhat akin to the biographical-data approach (see ch. 27), which has proved valuable in predicting air-crew success. Three of the four tests are commercially provided; one was constructed in the AAF program.

The Strong Vocational Interest Blank for Men, CE503A

Experimental administration of this well-known questionnaire was undertaken in an effort to estimate the validity of its large number of interest items, in predicting pilot success, with the prospect of development of a scoring key for pilots.

Description.—This commercially-provided (15) interest blank measures the degree of similarity between the expressed interests of the examinee and the professed interests of leaders in some 38 professions and occupations for which the test is scored. Ratings are made for such fields as artist, psychologist, physician, policeman, social science teacher, carpenter, ad[illegible] man, and for authors or journalists. Two non-occupational [illegible] are made; one for masculinity-femininity and the other for interest-maturity.

(1) *Internal characteristics.*—The blank consists of 400 items, each with three alternative responses. There are eight sections, which are listed in order:

Part I. *Occupations.* In this section the examinee indicates whether or not he would like each of some 100 different occupations, responding with "like," "indifferent," or "dislike." Sample occupations are: astronomer, governor of a State, and machinist.

Part II. *School subjects.*—The examinee indicates, by means of the same ratings used in section I, his interest in various school subjects. Sample subjects are: bookkeeping, military drill, and typewriting.

Part III. *Amusements.*—Using the same ratings as in the previous sections, the examinee records his first impression to various types of amusements, including sports, reading material, and places to visit. Samples are: roughhouse initiations, symphony concerts, and *Atlantic Monthly.*

Part IV. *Activities.*—In this section, the examinee again employs the ratings of "like," "indifferent," or "dislike," to indicate his interest in activities which range from the sedentary to the extremely active. Samples are: saving money, arguments, and pursuing bandits in sheriff's posse.

Part V. *Peculiarities of people.*—In this section, the examinee is instructed to record his first impression of various types of people, using the same ratings as in the preceding sections. Samples are: spendthrifts, people who talk about themselves, and people who don't believe in evolution.

Part VI. *Order of preference of activities.*—This section consists of 4 groups of 10 items each. The examinee indicates which 3 of the group of 10 items he likes most, and which 3 he dislikes most. The remaining four activities are checked "indifferent." Samples from 1 group of 10 items are: Operate (manipulate) the new machine; discover an improvement in the design of the machine; and determine the cost of operation of the machine.

Part VII. *Comparison of interest between two items.*—Pairs of items are given to which the examinee responds by checking one or the other of the pair of items, or if his preference is equal, he may indicate that choice. Samples are: Chauffeur or chef; Do a job yourself or delegate job to another; and Deal with things or deal with people.

Part VIII. *Rating of present abilities and characteristics.*—This section is divided into two parts. In the first, the examinee is to indicate by "yes," "?," or "no" whether or not each of a group of statements applies to him. Samples are: Usually get other people to do what I want done; Able to meet emergencies quickly and effectively; and Show firmness without being easy.

In the second part of this section, the examinee is to indicate which one of groups of three statements applies best to him. Samples are:

(1) Tell jokes well.	(2) Seldom tell jokes.	(3) Practically never tell jokes.
(1) Worry considerably about mistakes.	(2) Worry very little.	(3) Do not worry.

(2) *Administration.*—The test was administered without time limit, between 45 and 60 minutes being required for all examinees to finish.

(3) *Scoring.*—The author's keys were not used. It was planned, instead, to develop an empirical key for pilots. For purposes of cross-validation, answer sheets were split into two groups. Separate item analyses were accomplished for the two subsamples, and two scoring keys were devised. The criteria for scoring a response were: (*a*) a phi significant at or beyond the 5 percent level (0.11) and (*b*) a split of 85–15 or better. The evens group was scored with the odds key, and the odds group was scored with the evens key.

Statistical results.—Data are available for a group of approximately 650 pilots who took primary training, who were originally tested in June 1944 at Psychological Research Unit No. 1.

(1) *Item validity.*—After dividing the sample of answer sheets into two random halves, the responses to the items were correlated with the graduation-elimination criterion from primary pilot training. The distributions of the phi coefficients are shown in table 23.28. The phis are based on "yes" or "like" responses only.

TABLE 23.28.—*Item validation data for groups of pilots in primary training for the Strong Vocational Interest Blank, CE503A*

Phi	*f* (odds)[1]	*f* (evens)[2]
0.18 to 0.22	1	0
0.13 to 0.17	14	4
0.08 to 0.12	56	31
0.03 to 0.07	88	72
−0.02 to 0.02	117	115
−0.07 to −0.03	67	84
−0.12 to −0.08	15	49
−0.17 to −0.13	5	11
Total	363	366

[1] *N*=322. [2] *N*=325.

In interpreting these phi coefficients, it can be said that for an N of 322, a phi of 0.11 is significant at approximately the 5 percent level of confidence, and a phi of 0.15 is significant at the 1 percent level of confidence. In the odds sample 37 phis (25 positive and 12 negative) reached or exceeded the 5 percent level of significance, and 9 of these (8 positive and 1 negative) reached or exceeded the 1 percent level of significance. In the evens sample, 32 phis (10 positive and 22 negative) reached or exceeded the 5 percent level of significance, with 4 of these (4 negative) reaching or exceeding the 1 percent level of significance. There would seem to be an excess number of phis beyond the confidence limits, but, in view of the apparently unimodal distribution with its central tendency at zero, it is probable that there are few, if any, genuinely valid items in this collection for the prediction of primary pilot training success.

(2) *Cross-validation data.*—Cross-validation, nevertheless, was accomplished. Cross-validation data are presented in table 23.29. Since only one of the keys shows a validity that approximates the 5 percent level of significance (and then in the reverse direction), it is concluded that the scoring of this test with these empirical keys has no value in predicting pilot success in primary training. With the exception of this one key, the remaining biserial coefficients do not come close to significance.

TABLE 23.29.—*Cross-validation for a group of pilots in primary training, using a graduation-elimination criterion, for the Strong Vocational Interest Blank, CE503A*

Group	Key	Score	M_p	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
Odds[2]	Evens	Rights	25.21	27.57	6.50	−0.03	−0.01
		Wrongs	19.69	19.61	4.83	.01	.02
		R−W	5.52	5.96	10.18	−.03	−.01
Evens[3]	Odds	Rights	17.35	17.40	3.98	−.01	−.06
		Wrongs	12.04	11.21	3.29	[4].15	.20
		R−W	5.31	6.19	5.90	−.10	−.15

[1] Corrected to an unrestricted stanine standard deviation of 2.00.
[2] $N_t = 322$, $p_p = 0.70$, number of scored items = 85.
[3] $N_t = 325$, $p_p = 0.70$, number of scored items = 56.
[4] Significant at the 5 percent level.

Evaluation.—On the basis of the failure of the cross-validation results, it is concluded that this test cannot be used to predict success in primary pilot training.

It should be remembered that the attempt to derive an empirical key for this test was not similar to that customarily followed by Strong. His procedure would have compared experienced pilots' responses with those of other combined occupational groups. It would seem, however, that the procedure used in this study was more direct and should have yielded positive results if the items of the Blank are potentially discriminating for pilot selection.

Maller-Glaser Interest Values Inventory, CE514A

This inventory was administered in an attempt to determine the validity of personal-values scores in predicting success in flying training.

Description.—Allport and Vernon (1), following Spranger's classification in types of men (14), constructed a test to determine the relative prominence of six basic motives or evaluative attitudes that govern men's actions. These are: theoretical, economic, aesthetic, political, social, and religious. The Maller-Glaser simplification of the Allport-Vernon test is aimed at measuring four types of personal values: (*a*) social, (*b*) economic, (*c*) aesthetic, and (*d*) theoretical.

(1) *Internal characteristics.*—The inventory consists of 34 items, each with four alternative responses. It is divided into three parts.

Part I consists of items 1 through 10. Each item consists of four words. The examinee's task is to select the one word from among each group of four that pleases him most. A sample item is:

(a) Money.
(b) Research.
(c) Welfare.
(d) Masterpiece.

Part II includes items 11 through 20. It is concerned with word associations, in which the examinee is presented with a key word and four alternative responses. The task is to select the alternative which seems most closely associated with the key word. A sample is:

Civilization:
- (a) Justice.
- (b) Order.
- (c) Refinement.
- (d) Reason.

Part III includes items 21 through 34. The section is concerned with interests, in which the examinee indicates which of the four alternatives has most appeal for him. A sample item is:

If you were employed by a large automobile manufacturing concern and you had the necessary ability, which of the following positions would you prefer?
- (a) Handle the labor relations work.
- (b) Do research work on the development of a better automobile engine.
- (c) Direct a new market system for selling the car.
- (d) Work at improving the appearance of the automobile.

(2) *Administration.*—Sufficient time was allowed so that all could complete the test. Pertinent directions are:

This is a test of interests and values. Investigations have shown that men vary in the choices which they make in this test, and that these differences affect success in various types of activities. There are no right or wrong answers; simply indicate the one answer which appeals most to you.

(3) *Scoring.*—The authors' key was used in scoring. The test is scored for the four values of social (S), economic (E), aesthetic (A), and theoretical (T). One alternative for each item is scored for each value. Thus, there is an S, E, A, and T alternative to each item.

Statistical results.—Results are available for 524 pilots in primary training, originally tested in January 1944 at Psychological Research Unit No. 1.

(1) *Test reliability.*—Reliabilities were not computed.

(2) *Test validity.*—Validation results are presented in table 23.30.

TABLE 23.30.—*Validation data for pilots in primary training using the graduation-elimination criterion, for the Maller-Glaser Interest Values Inventory, CE514A*[1]

Score	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]
Social	6.64	6.80	3.06	−0.03	−0.09
Economic	13.27	12.11	4.55	[3].15	.02
Aesthetic	4.39	4.94	3.34	−.09	−.06
Theoretical	9.64	10.08	4.11	−.06	.08

[1] $N_t=524$, $p_p=0.80$.
[2] Corrected to an unrestricted stanine standard deviation of 2.00.
[3] Significant at the 5 percent level.

The rather large difference between the corrected and uncorrected biserial coefficients is due to the severe restriction of range in the stanine. In this sample, the standard deviation of the stanine was only 0.97.

(3) *Intercorrelations.*—The intercorrelations among the four categories are presented in table 23.31. The universal negative correlations are spurious and are due to the fact that the selection of one alternative automatically means rejection of others.

TABLE 23.31.—*Part-score intercorrelations for the Maller-Glaser Interest Values Inventory, CE514A, based on the scores of 524 pilots in primary training*

Score	S	E	A	T
S	...	−0.23	−0.27	−0.25
E	−0.23	...	−.40	−.61
A	−.27	−.40	...	−.15
T	−.25	−.61	−.15	...

Evaluation.—Inspection of the biserial coefficients for the Maller-Glaser Inventory reveals that only one category—economic—had a significant validity coefficient in predicting gradaution-elimination from primary pilot training. The exact nature of this relatonsihip is dubious since the corrections reduced this value nearly to zero. The three remaining coefficients were low.

This inventory does have certain advantages over certain other tests of its type. It is short, the type of items is diversified, and it has considerable interest for the examinee. It affords more freedom of choice, in that it provides four alternatives, than inventories of the yes-and-no answer variety.

The structure of the key may be open to question, in that each of the alternatives is scored for one of the four categories. Examinees frequently have conflicts between two equally strong choices. It might be well to call for two responses to each item. This would add to the information that can be extracted from the same items as well as minimize conflicts and reduce negative intercorrelations of scores.

Kuder Preference Record, CE515A

This preference blank was administered experimentally at Psychological Research Unit No. 1 in May 1944 in an attempt to validate its nine interest scores, several of which are not measured in the other personality inventories.

Description.—This test is intended for use primarily in the vocational and educational guidance of high-school and college students and in employee counselling (11). Measurements are made in nine general areas, which are listed below, together with some of the diverse activities which the author states are included in each area:

I. *Mechanical.*—Civil engineer, surgeon, industrial designer, fireman, and stonemason.

II. *Computational.*—Accountant, credit manager, bond salesman, purchasing agent, and traffic clerk.

III. *Scientific.*—Archeologist, oculist, agronomist, weather observer, radio operator, and science teacher.

IV. *Persuasive.*—Advertising manager, clergyman, psychiatrist, lawyer, receptionist, and retail merchant.

V. *Artistic.*—Art critic, furniture designer, illustrator, cabinet-maker, milliner, and photoengraver.

VI. *Literary.*—Copywriter, writer, advertising writer, reporter, English teacher, press agent.

VII. *Musical.*—Composer, accompanist, choir director, music teacher, singer, and sound engineer.

VIII. *Social Service.*—Camp director, psychologist, occupational therapist, sales manager, and policeman.

IX. *Clerical.*—Bookkeeper, billing machine operator, cost accountant,and court reporter.

(1) *Internal characteristics.*—This preference record consists of 168 items. Each item comprises a group of three activities, from which the examinee is to indicate the activity he likes the most, and that which he likes least, as in each of the following sample items:

Sample I

Visit an art gallery.
Browse in a library.
Visit a museum.

Sample II

Collect autographs.
Collect coins.
Collect butterflies.

(2) *Administration.*—The test is administered without time limit. Approximately 30 minutes are required for the majority of examinees to complete the test. Answers are recorded on the author's specially prepared answer sheets, suitable for machine scoring.

The directions include specific comment on the manner in which the answer sheet is to be accomplished and several sample problems. Pertinent directions are:

This blank is used for obtaining a systematic record of your preferences so that a picture can be obtained of how your preferences compare with those of others who have answered the questions. The blank is not a test of ability. There are no right or wrong answers. An answer is right only if it is a true expression of your preference * * * Some of the activities named in the following pages involve a certain amount of preparation and training. In such cases, make your choice on the assumption that you could first have the training and experience necessary for all the activities. Do not choose an activity merely because it is new or unusual. Make your choices on the basis of what you would like to do as a regular thing if you were equally familiar with all the activities.

In some cases you may find that you like all three activities in a group; in other cases you may find all three unpleasant. Please make your choices for every group even though the decisions may be hard to make.

(3) *Scoring.*—The scoring of the preference record is accomplished by means of the author's keys.

Statistical results.—Results are available for a group of 937 pilots who took primary training.

(1) *Test reliability.*—Reliability data are not available for this sample. The author of the record reports test-retest reliabilities, on a group of 41 graduate students, of 0.97, 0.98, 0.95, 0.96, 0.95, 0.95, 0.93, and 0.98, respectively, for the nine areas in the order given above.

(2) *Test validity.*—Validity data are presented in table 23.32 for a group of pilots in primary training.

TABLE 23.32.—*Validation data for pilots in primary training, using the graduation-elimination criterion, for the Kuder Preference Record, CE515A*[1]

Score	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]
Mechanical	86.13	85.48	15.57	0.02	0.11
Computational	33.12	33.39	9.35	−.02	−.02
Scientific	67.52	67.88	12.58	−.02	.07
Persuasive	68.33	68.47	16.34	−.01	−.08
Artistic	49.71	47.84	13.25	.08	.14
Literary	46.43	46.30	13.29	.01	−.03
Musical	19.13	18.40	8.96	.05	.08
Social science	63.15	65.67	14.27	[3]−.10	−.03
Clerical	46.48	46.25	12.10	.01	−.06

[1] N_t=937, p_g=0.77.
[2] Corrected to an unrestricted stanine standard deviation of 2.00.
[3] Significant at the 5 percent level.

It is noteworthy that only one of the biserial coefficients—that for the social-science score—is significant at the 5 percent level. The artistic score, which was expected to be low and perhaps negative on the bases of biographical-data and sports-and-hobbies test results, approached significance at the 5 percent level. Surprisingly, the mechanical score, which might be expected to have a high degree of validity for predicting pilot success, had little validity here.

(3) *Intercorrelations.*—The intercorrelations among the various interest scores are presented in table 23.33.

TABLE 23.33.—*Intercorrelations among part scores for the Kuder Preference Record, CE515A, for a sample of 937 classified pilots*

Interest	I	II	III	IV	V	VI	VII	VIII	IX
I. Mechanical	...	0.01	0.24	−0.27	0.20	−0.38	−0.26	−0.01	−0.07
II. Computational	0.01	...	.34	−.11	−.22	−.06	−.07	−.14	.51
III. Scientific	.24	.34	...	−.31	−.04	.17	−.23	−.06	−.05
IV. Persuasive	−.27	−.11	−.31	...	−.21	.15	.01	.01	.11
V. Artistic	.20	−.22	−.04	−.21	...	−.14	−.04	−.23	−.19
VI. Literary	−.38	−.06	.17	.15	−.14	...	.16	−.21	−.05
VII. Musical	−.26	−.07	−.23	.01	−.04	.16	...	−.14	.01
VIII. Social science	−.01	−.14	−.06	.01	−.23	−.21	−.14	...	−.25
IX. Clerical	−.07	.51	−.05	.11	−.19	−.05	.01	−.25	...

Upon inspection it may be seen that there is little or no degree of concomitant variation between the interest measures, with the exception of that between clerical and computational scores, which seems to be due to an overlapping in the type of questions in each area. The relationship between artistic and mechanical scores is just the reverse of what is expected on the basis of biographical-data results (see ch. 27). It is possible that the negative coefficients are spuriously high, due to the nature of the pairing of the various items throughout the test.

Evaluation.—On the basis of the obtained biserial coefficients, it is found that the Kuder Preference Record has no value in predicting success in primary pilot training. Three of the nine areas measured by the author's keys yield biserial coefficients different from those expected on the basis of other analyses. Thus the artistic and musical areas, which in other studies have yielded negative validity coefficients, in this case have yielded positive ones. The mechanical area, which almost without exception has had considerable validity for predicting pilot success, in this case was of negligible value. The chief explanation of this variation probably lies in the fact that these questions sample appreciation and interest in an area rather than experience, which is sampled by valid mechanical tests. The communality between mechanical interest, as measured by the record, and mechanical experience must, therefore, be very low. The evidence can be taken to mean that the mechanical-experience factor is properly named; at least, that it is not an interest factor.

Teacher Preference Scale, CE426A [a]

This test was developed for the purpose of assessing several hypothesized personality characteristics. It was expected that the examinee would reveal his personality by indicating the type of teacher he prefers. In using a teacher-preference paired-activity scale, it was hoped that the underlying principle of the test would be hidden sufficiently so that students would not be able to detect its true purpose.

Description.—The hypothesized traits for which the scale is scored are:

(*a*) Excessive demand for definiteness of structure (SD): Individuals vary in the degree to which they require definiteness in the training situation. Some may tolerate more unknown elements than others. The student with a low tolerance for the uncertain reacts to the training situation with hesitancy, has difficulty in building and maintaining confidence, and is hypersensitive to change. His behavior is characterized by confusion, tenseness, blocking, and anxieties.

(*b*) Decision difficulties (DD): In some cases the failure of a student lies in his reactions to situations that require him to arrive at decisions quickly and appropriately. In such situations, alternatives must be kept in mind, weighted, integrated, and a practical decision reached. The individual with decision difficulties may know the various factors involved, yet not arrive at decisions quickly and effectively. The net result usually is undesirable mechanical flying.

(*c*) Ego-sensitivity (ES): This category refers to the individual who possesses strong, unsatisfied ego needs that are sufficiently dominant to

[a] Developed at Psychological Research Unit No. 3. Chief contributor: Lt. Jacob S. Kounin.

interfere with his progress and performance. Included here are such traits as hypersensitivity to criticism, extreme desire for independence, and self-consciousness. The attention to self tends to leave less attention for other things pertaining to flying, makes the person hesitant to secure information, results in tension, hesitancy, and worry about failure to too great an extent.

(*d*) Social-sensitivity (SS): Somewhat related to ego-sensitivity, this category of social-sensitivity includes such traits as over-attentiveness to the instructor as a person and over-dependence upon the instructor.

(*e*) Normal (N): A normal key includes those responses not keyed adversely for any of the other traits.

(1) *Internal characteristics.*—The scale is divided into two parts. In part I the examinee indicates the type of teacher that he would like to have as an instructor. Pairs of characteristics of teachers are presented, and the examinee chooses the preferred characteristic in each item. In part II the examinee indicates the type of teacher that he would not like to have as an instructor, again by choosing between pairs of characteristics. Sample items follow:

Sample items, part I: (Choose type of teacher most preferred)

1. A. The teacher who is a "square shooter" and a "regular fellow."
   B. The teacher who is able to gather and judge facts and arrive at clear conclusions.
2. A. The teacher who pays special attention to the slow or maladjusted pupil.
   B. The teacher who makes work interesting by using examples and illustrative material.

Sample items, part II: (Choose type of teacher most disliked)

1. A. The teacher who is dishonest.
   B. The teacher who is changeable and inconsistent.
2. A. The teacher who always hesitates and never seems to make up his mind.
   B. The teacher who stands for a lot of foolishness and waste of time.

(2) *Administration.*—The test is group-administered, 23 minutes being allowed for each part of the scale. There are 80 items in each part. At the end of 10 minutes, the examinees are told that they should be half-way through the part. The time limit was set so that at least 80 percent of the group would complete the test.

(3) *Scoring.*—The two parts are scored separately according to the five a priori categories: Normal, ego sensitivity, social sensitivity, structure demands, and decision difficulty.

Statistical results.—Data on internal consistency and score intercorrelations are available for pilots in primary training originally tested at Psychological Research Unit No. 3 with validation data available for part I only.

(1) *Test validity.*—Validity data are available for a group of pilots in primary training, using the graduation-elimination crterion. These results are presented in table 23.34.

TABLE 23.34.— *Validation data for a group of pilots in primary training, using the graduation-elimination criterion, for part I of the Teacher Preference Scale, CE426A*[1]

Score	M_p	M_e	SD_t	r_{bis}	r_{bis}[2]
N	35.61	36.19	4.45	−0.08	−0.12
ES	9.02	8.91	2.04	.03	.04
SS	11.81	12.35	2.27	[3]−.14	−.16
SD	10.74	10.66	2.12	.02	.04
DD	9.49	9.71	1.87	−.07	−.05

[1] N_t=422, p_e=0.80.
[2] Corrected to an unrestricted stanine standard deviation of 2.00.
[3] Significant at the 5 percent level.

Only the negative relationship between the social sensitivity score and success in primary pilot training is significant at the 5 percent level.

(2) *Intercorrelations.*—Score intercorrelations were computed for part I, part II, and between corresponding scores in parts I and II. The correlations between "normal" scores and others are, of course, spurious, because of the scoring method. These results are presented in tables 23.35, 23.36, and 23.37. On the basis of the intercorrelations between the two parts, a low degree of communality is indicated. Thus, it would have been advisable to validate each part of the test rather than to base conclusions on part I as symptomatic of both parts.

TABLE 23.35.— *Intercorrelations of scores for part I of the Teacher Preference Scale, CE426A*[1]

Key	N	ES	SS	SD	DD
N	...	−0.32	−0.25	−0.32	−0.27
ES	−0.32	...	.11	.03	.12
SS	−.25	.11	...	.09	.13
SD	−.32	.03	.09	...	.14
DD	−.27	.12	.13	.14	...

[1] N_t=422 pilots in primary training, originally tested from September 13 to Nov. 21, 1944.

TABLE 23.36.— *Intercorrelations of scores for part II of the Teacher Preference Scale, CE426A*[1]

Key	N	ES	SS	SD	DD
N	...	−0.75	−0.59	−0.68	−0.66
ES	−0.75	...	.45	.34	.28
SS	−.59	.45	...	.19	.15
SD	−.68	.34	.19	...	.50
DD	−.66	.28	.15	.50	...

[1] N_t=273 pilots in primary training in class 44B, originally tested in November 1943.

TABLE 23.37.— *Intercorrelations of scores for part I with those of part II for Teacher Preference Scale, CE426A*[1]

Key	r_{tt}	r_{tt}
N	0.39	0.56
ES	.31	.47
SS	.25	.41
SD	.27	.42
DD	.35	.52

[1] N_t=273 pilots in primary training in class 44B.

Evaluation.—Inspection of the data for the Teacher Preference Scale, part I, indicates that only one category, social sensitivity, has any promise of validity. The coefficient is negative in sign, indicating that students with low social sensitivity scores have a slight advantage in graduation from primary training. The relatively low intercorrelations between the scores on part I and those on part II indicate unacceptably low reliabilities. They may indicate a functional difference between the two ways of phrasing the questions.

Two criticisms have been made of the rationale of this test. First, it seems likely that, if this test were administered to examinees in any particular phase of training, they might have certain instructor or teacher stereotypes which would interfere with the validities of their answers for prediction in some other stage of training. Secondly, it seems questionable whether an examinee's "ego needs" can be projected clearly, simply by stating his preferences for teachers.

Another objection lies in the fact that the examinee is forced to choose between two alternatives, without any recourse to modified decisions. Thus, lack of a third alternative, such as "don't know" or "haven't experienced this," might well lower the discrimination required of the examinee. Group-test administrators report that, having only two possibilities from which to choose, the examinee tends to adopt a somewhat superficial attitude toward his choices. Apparently, a few questions which force the examinee to choose between what, for him, are situations beyond the scope of his information or experience, or which do not fit him, will adversely color the nature of his interest and effort toward the other items in the test.

This type of scale, if carefully developed, might well serve as a selective device in indicating the type of instructor (ground school or flying) needed by each student. Likewise, a complementing selective device applied to instructors might well yield information which could result in a matching of student and instructor, which would yield an optimal teacher-student relationship. The test apparently has little promise as a pilot-selection instrument.

SUMMARY AND EVALUATION

This chapter has demonstrated that personal inventories and preference inventories generally fail in the two primary objectives that were set for assessment. First, with two, or possibly three, exceptions, none was able successfully to predict graduation from primary pilot training at significant levels of confidence. Second, item-validation studies generally failed to yield many items with statistically significant validities. It may be that with a higher standard of item selection—at the 1 percent level of confidence or at the 5 percent level in both odds and evens samples—better success would have been attained. Even with this highly restricting standard, a sufficiently large number of items could be pooled from several inventories to make further study with them profitable.

A very useful purpose has been achieved, however, in the general failure of these tests to predict air-crew training success. It serves as confirmation of the belief that temperament tests must be constructed specifically for the job intended. It is obvious from the data obtained that almost all these tests, not constructed for the task at hand, fail to validate successfully against a training criterion.

In the discussion of the failure of this type of test to validate successfully against a training criterion, it must be remembered that they may be of value in predicting success or failure in combat in terms of combat neuroses. The chief difficulty here is in obtaining actual combat validation data.

It is possible, also, that even in training, the graduation-elimination criterion or some other job-proficiency criterion is not a suitable one for the validation of temperament tests. Many a person with poor temperamental traits may by extra effort and under external pressure show satisfactory job proficiency. The criterion might better be in the areas of trainee satisfactions and adjustments and of instructor and supervisor satisfactions with him as a person with whom they must deal. Such criteria might yield quite different validities than those found for tests described in this chapter.

Criticisms of the tests have been made both from the standpoint of the examinee and from that of the administrator. The length of the majority of the tests is too great, even if moderate valse were received. Examinees have complained of actual boredom by the time they have completed some of the longer inventories; some have insisted that there is actual repetition of items.

The structuring of the items is another source of difficulty. In some cases, fear, pride, or shame, may cause an examinee to falsify his answers. Knowledge—or in some cases assumed knowledge, which is often faulty—as to what is desired in the classification schema colors many responses, at times causing the examinees to attempt to outguess or outwith the purpose of the inventory. The use of inventories that provide only "yes" and "no" responses, without recourse to a third category of indecision of "?," has been stated by examinees to be frustrating and to affect their reactions to the remaining items in the inventory.

Two suggestions have been made for further study:

(1) Tests such as the Teacher Preference Scale, CE426A, if carefully developed, might well serve as a selective device in indicating the type of instructor needed by each student. A complementing technique applied to instructors might yield information which could result in a matching of student and instructor which would afford an optimal teacher-student relationship.

(2) A more comprehensive measuring device than that of the individual inventories would be afforded by combining the scores from several complementary inventories—such as the Guilford-Martin Personnel

Inventory, the Inventory of Factors S T D C R, and G A M I N Inventory. In this instance, the scores for the 13 factors might be plotted on a composite graph, by means of which significant profile configurations of traits would be revealed. Validation of such configurations would prove an interesting and possibly fruitful investigation.

BIBLIOGRAPHY

(1) Allport, G. W., and Vernon, P. E. *A Study of Values*, Boston: Houghton-Mifflin, 1931.

(2) Bernreuter, R. G. *The Personality Inventory*, Stanford University: Stanford University Press, 1935.

(3) Darley, J. G., and McNamara, W. J. *Minnesota Personality Scale for Men*, New York: The Psychological Corporation, 1941.

(4) Guilford, J. P., An Inventory of Factors S T D C R, *Manual of Directions and Norms*, Beverly Hills: Sheridan Supply Co.

(5) Guilford, J. P. A simple scoring weight for test items and its reliability, *Psychometrika*, 1941, pp. 367-374.

(6) Guilford, J. P., and Martin, H. G. The Guilford-Martin Inventory of Factors G A M I N, *Manual of Directions*, Beverly Hills: Sheridan Supply Co., 1943.

(7) Guilford, J. P., and Martin, H. G. The Guilford-Martin Personnel Inventory, *Manual of Directions*, Beverly Hills: Sheridan Supply Co.

(8) Hathaway, S. R., and McKinley, J. C. *Manual for the Minnesota Multiphasic Personality Inventory*, Minneapolis: University of Minnesota Press, 1943.

(9) Humm, D. G., and Wadsworth, G. W., Jr. The Humm-Wadsworth Temperament Scale, *Amer. Jr. Psychiat.*, 92, 1, July 1935, pp. 163–200.

(10) Humm, D. G., and Wadsworth, G. W., Jr. The Humm-Wadsworth Temperament Scale, *Manual of Directions*, Los Angeles: D. G. Humm Personnel Service, 1942.

(11) Kuder, G. F. *Manual for the Kuder Preference Record*, Chicago: Science Research Associates, 1943.

(12) Rosanoff, A. J. *Manual of Psychiatry*, Sixth Ed., New York: John Wiley and Sons, 1927.

(13) Shaffer, L. F. *Psychology of Adjustment*, New York: Houghton Mifflin, 1936, pp. 320-323.

(14) Spranger, E. *Types of Men* (translated from the fifth German edition of *Lebensformen*), New York: G. E. Stechert, 1928.

(15) Strong, E. K., Jr. *Vocational Interest Blank for Men* (Revised), Stanford University: Stanford University Press, 1938.

CHAPTER TWENTY-FOUR

Clinical Type Procedures[1]

INTRODUCTION

The introduction of clinical types of procedures in the air-crew classification program constitutes a sharp departure from standard written tests both in approach and in techniques. For this reason it is appropriate to state briefly, by way of preamble to the procedures themselves, the general rationale underlying this use.

Fairly specific aims were set forth in the project proposal at Psychological Research Unit No. 1 for experimental application of clinical procedures to air-crew classification: (1) To determine the prognostic efficiency of various clinical procedures, separately and in combination; (2) to discover and interpret interrelationships among clinical techniques and other measurements; (3) to provide leads for the development of tests that could be empirically validated; and (4) to provide case studies of the personal characteristics of examinees from which could be built up eventually a man analysis (in terms of personality characteristics) of performance in pilot training and, if possible, in combat.

The clinical-type procedures attempt to emphasize the interaction among personality traits within the individual. Consideration of the role of compensating or balancing factors is fundamental in these procedures in evaluating any single characteristic of the individual. Because a global approach is basic to these techniques, they afford the possibility of investigating complex interrelationships of personality that are difficult to ascertain through the use of other methods.

Many of the tests to be described in this chapter are of the type that in ordinary clinical practice are individually administered, such as the Rorschach *Psychodiagnostik* and the Thematic Apperception Test. In an attempt to adjust them to meet a large testing load, however, either the scoring or administration, or both, were in some trials adapted to a large-scale method of handling. Such devices—as a special scoring system for the individually-administered Rorschach, a group-administration method for the Rorschach and for the Thematic Apperception test—represent attempts to streamline individual clinical-type procedures to fit a large-scale testing program.

[1] Written by S/Sgt. Arthur Z. Cerf.

The Clinical Techniques Project

A clinical-procedures group[2] that sought to explore the possibilities of utilizing a holistic approach to the subject of classification for air-crew positions was established at Psychological Research Unit No. 1. At one time the objective of this group was to establish an entirely separate stanine, based on the results of clinical procedures, which was to be validated against the criterion of graduation-elimination from primary-pilot training. This goal was never realized.

The complete battery established for the clinical-techniques project cuts across the lines of several chapters in this volume. The tests are:

(1) Projective techniques:
- (*a*) The Rorschach Test, CE701A.
- (*b*) Thematic Apperception Test, CE706A.

(2) Observational techniques:
- (*a*) Observational Stress Technique, CE710A.
- (*b*) Observation During Psychomotor Testing Rest Period, CE709A.
- (*c*) The Interaction Test, CE425A.
- (*d*) Observation of Atypical Behavior During Psychomotor Testing, CE708A.
- (*e*) Conference on Occupational Background and Interpretation of Test Score, CE707A.

(3) Printed tests:
- (*a*) The Behavior Preference Questionnaire, CE432A.
- (*b*) The Personal Audit, CE431A.
- (*c*) Occupational Experience Blank, CE603A.

(4) Self-rating techniques:
- (*a*) Indices of Self-Confidence, CE427A.

Plan of Approach

The content of this chapter is in two sections. The first is concerned with projective techniques within the accepted definition of the term, of which probably the best known are the Rorschach and the Thematic Apperception tests. These projective instruments range from tests that are individually administered and scored through others, such as the Empathetic Response test, that are group-administered and machine-scored.

In the second section, observational techniques are presented. These techniques involve the observation and rating of the behavior of examinees under stress situations either in the performance of actual apparatus tests, such as the observation of atypical behavior during psychomotor tests, or in tests specifically designed for observational procedures such as the Observational Stress test.

[2] Members: Lt. Avrum. H. Ben-Avi, Sgt. Gerald S. Blum, Lt. Mason Haire, Lt. John S. Harding, Lt. George S. Klein, Lt. John W. Nygard, Sgt. Harold M. Proshansky, Lt. John W. Rothney, Sgt. Leo Srole, Sgt. Bernard Steinzor, Capt. Donald E. Super, Staff/Sgt. John L. Wallen.

Except where noted to the contrary, the data that follow are based upon unclassified aviation students tested at Psychological Research Unit No. 1 during the period May 27 to June 31, 1943. Those who entered pilot training were in classes 44C, 44D, and 44E.

PROJECTIVE METHODS

Projective methods are intended principally as probing instruments, going beyond the conscious or superficial responses of individuals in determining personality content and structure. In this, they promise types of information regarding goals, values, conflicts, anxieties, and emotional complexes beyond that obtainable by the usual inventory procedures. The promise, further, of a total-personality picture in which is revealed not simply isolated traits and attitudes but their interrelationships in the structure of personality is one not afforded by other approaches.

In method and stimulus material, the projective techniques depart from the questionnaire type of personality test. The method is founded, essentially, on the assumption that the individual reveals his ways of organizing experiences when he is presented with stimuli that are relatively poorly structured as to meaning and that he himself organizes through the projection of meanings, ideas, and feelings. The projective methods thus allow for greater freedom and spontaneity in response than do standard ratings and questionnaires; responses are less biased by prearrangement of test items, as in psychometric tests, and do not depend on self diagnosis or introspective ability.

The Rorschach Test, CE701A

It was thought that further information on the aviation trainee's intellectual, emotional, and motivational characteristics, and the patterning of these traits, beyond that revealed by printed tests, would reveal significant information in the prediction of air-crew success. Accordingly, experimentation with a standardized administration of the Rorschach technique (4) was undertaken.[3]

Description.—The Rorschach method is essentially a procedure for revealing the personality of the individual as an individual, as contrasted with rating or assessing him in terms of his likeness or conformity to social norms of action and speech. It is just because a subject is not aware of what he is telling and has no cultural norms behind which to hide himself that the Rorschach and other projective methods are so revealing (1).

In contrast to this approach to personality diagnosis, personality inventories and questionnaires attempt to establish more rigidly controlled and standardized situations. In the questionnaire method, however, the examiner is deprived of the possibility of understanding how and why

[3] Administration and interpretation was executed chiefly by the following individuals, most of whom are members of the Rorschach Institute: Cpl. James A. Christenson, Pvt. Herman Feifel, Sgt. Harold M. Proshansky, Cpl. Walter J. Reis, Sgt. Bernard Steinzor, and Lt. Herbert J. Zucker.

the examinee arrives at a particular result. There is a second important disadvantage in the use of questionnaires, in that the examiner must rely almost entirely on information which the subject is willing and able to furnish. The Rorschach method seeks to overcome these difficulties.

(1) *Internal characteristics.*—In the typical application of the task, the examinee is presented with a standard series of 10 ink-blot pictures, reproduced on cards 7 inches by 9½ inches (8). In five of the blots, colored ink is used in addition to black ink. The examinee is asked merely to report what he sees in the blots. The blots themselves, while symmetrical, have little structure. It is assumed that the examinee, by the very ambiguity of the blots, must himself organize them in order to give them some form or meaning. In the main part of the test the examinee gives his free responses to each blot, the examiner offering no further encouragement after his initial instructions, merely recording verbatim the examinee's responses.

The inquiry follows the main part of the test. After the examinee has given his responses to all of the 10 cards, the examiner returns to each card, reads the responses, and inquires into the way in which the responses were formed; whether, for example, it was shading or color, and whether the whole or a part of the blot determined the response. From an analysis of the frequencies of responses within the various categories, and through qualitative considerations of the types of responses, a picture of the personality of an individual is constructed from the projected material, including both intellectual and nonintellectual aspects.

(2) *Administration.*—In the large-scale administration of the test, the gathering of responses to the ink-blot cards, the scoring and the interpretation of the records were, in a sense, independent steps. The same examiners did not, in all cases, participate in all three phases on the same records. The steps in the processing of a record were:

(*a*) Recording of responses to the cards. Some preliminary scoring was done at this time.

(*b*) The examiner's first clinical impression formed entirely from the behavior of the examinee in the test situation. Solely on the basis of this over-all impression, the examiner made a clinical prediction of success in primary pilot training (CE701A–I). The prime consideration underlying these predictions was the examiner's knowledge of the results of job analyses for pilots.

(*c*) Initial scoring of the record by the examiner.

(*d*) Checking of the scoring.

(*e*) Tabulation and summary of the scores.

(*f*) Final interpretation of the record.

(*g*) Clinical prediction of success or failure in primary pilot school (CE701A–II), based on scores and interpretation of the record.

One examiner tested one examinee at a time. Each examiner tested approximately four aviation students during each testing day. Most examiners were members of the Rorschach Institute. A rigorous course in interpretation was conducted for all examiners.

A few minutes were spent during the first part of the testing period in informal conversation with the examinee to put him at ease. Then the test was begun with these instructions:

> This is a test of visualization. It is somewhat different from others you have taken, and you'll probably find it a more interesting experience. It consists of 10 cards made up out of ink blots. You know you can drop ink on a sheet of paper, fold it, smear it over, and when you open it, find a "blotto." Different people see these cards in different ways. We would like to know what you see in these cards. There are no right or wrong answers. You can say as little or as much as you want about each card. Time does not matter in this test. When you've said all you want about a card, place it face down on the table, and I will give you the next one. Here is the first card * * * What might this be?

Upon completion of the 10 cards, the examiner began the inquiry with these instructions:

> That completes the first part of the test. Here are the cards. In order to score your responses adequately, it is important for me to be sure how and why as well as where you saw the particular things you did, so I can see them exactly as you did.

There were no time limits. Average testing time ranged between 75 and 120 minutes.

(3) *Scoring.*—Two general scoring treatments were accorded the data:

(*a*) CE701A-I.—Clinical predictions of success in primary pilot training were based on the over-all clinical impression. This rating was made in most cases immediately after the administration of the test. Ratings of the examinees' self-confidence, on a five-point scale, were also made at this time.

(*b*) CE701A-II.—Clinical predictions of success in primary pilot training based on interpretations of the Rorschach responses. These ratings were made after the records had been scored and interpreted.

It is now in order to list the interpretive significance of each of the various scoring categories or signs for air-crew training. This interpretative schema reflects as much as possible the opinions held in common by the examiners concerning the requirements of air-crew duties. If a certain characteristic of a category score is regarded as a positive or a negative indicator, it should be clear that this evaluation is made in relation to the prospect of success in pilot training; it does not represent a value judgment of an examinee. It is obvious, also, that these positive and negative weightings do not follow necessarily from the Rorschach interpretive procedure. A brief interpretation of the categories as agreed upon by the clinical procedures group follows:

1. *Inner life.*—*a.* M.—The absence or presence of M (human movement) was given no special significance except in cases where many M's might compensate for unfavorable features of the record. The pres-

ence of some human movement, two or more M's, was of positive interpretive value as a compensating factor.

b. FM.—The lack of FM (animal movement) was viewed as a negative indicator (i. e., associated with failure). A well-developed FM column, considered in conjunction with the characteristics of the content of the responses themselves, numerous responses, and an adequate number of whole responses (W's), indicates in the Rorschach schema an adequate fund of energy. These characteristics were taken to mean in the present study that the person had generally desirable drive and, specifically, a real desire to fly in the air-crew situation.

c. m.—If the number of m (inanimate movement) responses outweighed both M and FM, it was considered a poor sign, since it indicated intrapersonal conflict. A few m's, coupled with the presence of anxiety and an inability to make an adequate adjustment, indicated for the interpreters that the individual was a very poor prospect for pilot training.

2. *Outer life.*—*a*. Sum C.—The absence or marked suppression of sum C (color) responses was given considerable negative weight. The CF response (determined by both color and form, but primarily color) seemed to be more characteristic of the group than FC preponderant, and this reaction was presumed to be a favorable sign, even if it was in the FC column. If the dominant color responses were explosive, such as volcano or fire, the examinee revealed poor self control and was considered a poor prospect. The presence of FC responses alone was regarded as slightly unfavorable, since it indicated a too careful individual who was not very spontaneous. Pure C responses, which were very few in number, had to be considered in conjunction with other factors before any significance could be attached to them. In general, the gross reactivity to color as expressed in sum C was regarded as the more important factor. If the individual evidenced a basic extroverted pattern (as revealed by the fact that he offered many responses to the cards containing color) but did not utilize the color itself in his responses, the absence of color responses was regarded as especially serious.

3. *Control.*—Poor control was assumed to be a negative element in prediction when F percent (the percentage of responses which are determined exclusively or primarily by form) was below 20. This could be compensated for by indications of control in other responses, such as good form perception in M, FM, and color area. If the student had more than 60 percent F, he was considered to be constricted. In terms of flying, this meant presumptively that he was unable to shift his attention with enough rapidity and that he lacked flexibility in his approach to situations. This constriction was regarded as especially serious if it resulted in a repression of the inner life (in the case of an individual who was characteristically an introvert) or in a repression of the outer life (in the case of a basically extroverted person).

4. *Sensitivity.*—Fc, c.—The use of texture (*c*) in responding was considered a positive sign. It was taken to indicate a degree of sensitivity and tact which would stand the student in good stead in his relations with instructors and his fellows. It also indicated a certain plasticity and an affinity with objects. Thus, an individual possessing these characteristics might be able to feel his way through a maneuver and fly by the seat of his pants. The use of Fc (response concept of definite form and texture) indicated a more elaborate way of indicating control. An important qualification must be stated, however. If Fc plus c outweighed the F column or sum C in an extroverted personality, this syndrome was negatively weighted. In general, also it was felt, as in the case of F, that too high a frequency was as unfavorable as too low a frequency.

5. C'.—Only a small proportion of cadets utilized black, white, or grey as surface color. This was, therefore, not regarded as a very important sign except when the C' responses were more frequent than the color responses, thereby indicating a tendency to avoid full responsibility or involvement with social and emotional experiences.

6. *Anxiety.*—k (toned-down shading effects), K (use of diffusion), FK.—These determinants did not appear frequently. Usually the existence of anxiety had to be derived from other aspects of the record. But if there were many k and K responses, the diagnosis of anxiety was more certain. If the chiaroscuro responses outweighed the development of the inner life, M and FM, the picture was regarded as unfavorable. Presence of FK, indicating an introspectiveness and tendency toward self-analysis, somewhat diminished the seriousness of signs of anxiety. If FK outweighed F, however, the examinee was regarded as too self-conscious for purposes of successful performance in the training situation.

7. *Mental approach*—W, D, d, Dd+S.—The normal percentage in the use of location (as defined by Klopfer and Kelley [4]) was taken as an important sign for success in air-crew training. A well-balanced and elastic mental approach was believed to be one of the most desirable assets of the trainee. If the individual had an unusually high W percent (percentage of whole responses), this indicated too great a preoccupation with a phantasy life or a consuming intellectual ambition, and these were taken as unfavorable signs. Too few W's indicated an inability to integrate and organize a task or situation. A low W percent usually implied a high D percent (percentage of responses utilizing large usual details). This indicated too great an emphasis on everyday facts and details. In general, however, this was not assigned great importance, and the absence of D was considered significant only when there was no further attempt at analysis within the whole response. Dd (unusual detail) +S (space), and *d* (small usual detail) were also assigned a negative weight when they far exceeded the norms defined by Klopfer and Kelley.

8. *Erlebnistyp ratios.*—(*a*) M : Sum C; (*b*) (FM+m) : (Fc+c+a'); (*c*) (8+9+10) percent. These three ratios were mainly considered for

703320—47—41

their consistency. It was assumed that an extraverted individual was more suited for pilot training than an introvert. However, the introvert-extravert tendency was not regarded as important as whether the individual was living out his basic tendencies. In other words, the question was whether he was "being himself." Both an introversive tendency with no M's or FM's and an extraverted picture with no color reactions were negative indicators. The general assumption, then, was that an individual will not function at his optimum efficiency if he lives against his basic tendencies.

9. *A percent (percentage of animal responses).*—This was not regarded as a very important factor. Only if A percent exceeded 60 percent and was regarded together with other indicators of mental stereotypy (such as a large number of popular responses and little variety of response) was it allowed some negative significance.

10. *P—O (popular minus original responses).*—This was practically never used as a datum for a prediction.

11. *Frequency of responses and rejection of cards.*—The number of responses never received more weight than the quality of responses. It was soon observed, however, that numerous examinees gave only 10 responses. This posed a rather difficult problem for interpretation. Not only were the tabulated statistics rendered less reliable, but the quantity of material in the record was often very meager. Predictions in these cases were consequently made without much confidence, and the examiner's clinical impression was relied upon more heavily than usual. In general, it was felt that the more responses a person gave, the better were his chances of passing training. If the individual gave 20 or more responses (R's), however, and yet revealed constriction or repression of his functions to an unusual degree, his chances of success were considered to be decreased. There were few individuals who rejected cards (failed to give any response at all). Those who could not find an answer to one or more cards were considered as poor training prospects.

12. *Time.*—The time taken by the examinee to respond to a card was very limited as a source of interpretation. Only when a student was very slow in responding, or when his reaction time to any particular card indicated color or shading shock, did the time contribute to the interpreter's judgment.

Statistical results. (1) *Examiner differences.*—The influence of the examiner upon the number of responses was studied. The means and standard deviations of numbers of responses are presented by examiner in table 24.1. The critical ratios for differences among examiners are given in table 24.2. Twelve differences are significant at the 1 percent level of confidence and three more at the 5 percent level—a total number that would appear to be well above the expectation on the assumption of homogeneity of examiners.

TABLE 24.1.—*Distribution statistics on number of responses obtained by nine examiners on the Rorschach Test, CE701A*

Examiner No.	N	M	SD	Rank order
I	75	23.2	15.4	2
II	49	17.5	9.8	7
III	67	21.5	12.1	4
IV	66	15.1	7.2	8
V	72	19.9	10.4	6
VI	32	24.3	16.0	1
VII	44	14.6	6.5	9
VIII	26	22.7	10.9	3
IX	66	20.5	14.4	5

TABLE 24.2.—*Critical ratios between means of responses for nine examiners on the Rorschach Test, CE701A*

Examiner No.	II	III	IV	V	VI	VII	VIII	IX
I	[1]2.54	0.74	[2]4.11	1.57	0.31	[2]4.23	0.21	1.09
II	......	[1]1.98	1.45	1.25	[1]2.15	1.66	1.99	1.33
III	......	......	[2]3.72	.81	.86	[2]3.84	.43	.44
IV	......	......	......	[2]3.15	[2]3.11	.34	[2]3.22	[2]2.72
V	......	......	......	......	1.44	[2]3.30	1.12	.30
VI	......	......	......	......	......	[2]3.22	.45	1.13
VII	......	......	......	......	......	......	[2]3.34	1.76
VIII	......	......	......	......	......	......	......	[2]2.87

[1] Significant at the 5 percent level.
[2] Significant at the 1 percent level.

(2) *Temporal differences.*—The influence of time of examination upon numbers of responses was also studied. The results are presented in table 24.3.

TABLE 24.3.—*Mean numbers of responses obtained at different times of the day for the Rorschach Test, CE701A*[1]

	Morning	Midday	Afternoon
Mean	18.3	19.3	20.5
SD	12.6	10.9	13.2

[1] Number of cases=497.

(3) *Reliability.*—Test reliability was not estimated.

(4) *Validity of clinical impressions.*—Validity results are available against the criterion of graduation or elimination from primary training. In table 24.4 are shown the validity data for the over-all clinical impressions and predictions, made immediately after the administration of the test (CE701A–I).

TABLE 24.4.—*Validation data for clinical predictions based on Over-all Clinical Impressions (CE701A-I), based upon the Rorschach Test, CE701A-I, for groups of pilots in primary training, class 44C, using the graduation-elimination criterion*

N_t	p_t	r_{bis}
293	0.89	0.12
190	.81	.06

Validation of clinical predictions of success in primary pilot training, based upon interpretations of the scored Rorschach records (CE701A-II), however, was more promising. For a group of 281 pilots in primary training, 92 percent of whom graduated, the biserial coefficient of correlation was 0.23, which is significant beyond the 5 percent level of confidence. The corrected biserial was 0.26. Because of the marked differences between examiners, however, it is felt that these results are not definitive.

(5) *Validity of single categories.*—Individual scoring categories, 25 in number, were validated for two groups of pilots in primary training in order to evaluate some of the noninterpretive or direct measures yielded by the Rorschach technique. Each record was scored with respect to the location, determinant, and content of responses, according to the system described by Klopfer and Kelley (4). The data are presented in table 24.5. For group I, a biserial correlation of 0.19 is required for significance at the 5 percent level and of 0.25 at the 1 percent level. For group II, the required coefficients are 0.21 and 0.27. The biserial correlations are seen to be generally low, and those that seemed significant in group I proved to be of doubtful significance in the revalidation of the categories in group II.

TABLE 24.5.—*Validation data for 25 single Rorschach categories for the Rorschach Test, CE701A, graduation-elimination criterion for two samples of pilots in primary training*

Category	M_g		M_e		SD_t		r_{bis}	
	I[1]	II[2]	I	II	I	II	I	II
R	18.49	20.90	15.75	19.15	10.94	12.65	0.14	0.08
T	738.00	758.00	670.00	700.00	450.00	448.00	.08	.07
T/R	44.77	43.00	40.31	44.00	22.81	20.26	.10	−.03
R_{TA}	21.12	22.00	20.32	23.00	13.55	17.0	.03	−.02
T_{TO}	27.21	28.00	24.69	24.00	[illegible]	22.0	.08	.14
W	9.17	8.80	7.31	8.65	4.09	17.35	.24	.02
W percent	60.16	53.26	55.75	55.3[illegible]	27.85	27.94	.08	−.04
D	7.06	8.38	6.72	6.73	6.70	6.95	.03	.13
D percent	31.67	34.29	37.63	35.0	20.47	20.06	−.15	−.03
Dd+S percent	8.0	9.39	6.0	8.89	7.99	10.30	.13	.03
M	1.43	1.75	.91	1.51	1.78	1.77	.15	.08
FM	3.84	3.89	3.03	4.05	2.72	2.78	.15	−.03
F	6.69	7.42	5.81	6.73	5.91	7.16	.08	.06
F percent	34.30	32.70	35.95	31.45	17.04	16.65	−.05	.04
FK+F+Fc percent	39.61	43.76	40.31	41.59	17.37	16.20	−.02	.08
Fc	1.43	1.65	1.09	1.70	1.71	1.95	.10	−.01
Fc+c	1.79	2.20	1.50	2.03	1.85	2.35	.08	.04
FK+K+k	1.03	1.42	1.28	1.16	1.34	1.68	−.08	.09
Sum C	2.36	2.20	2.06	1.84	3.50	1.96	.04	.11
FC	1.40	1.67	1.13	1.46	1.55	1.48	.09	.08
CF	1.55	1.56	1.41	1.16	1.46	1.58	.05	.14
8, 9, 10 percent	36.00	35.71	37.94	35.97	9.05	8.98	−.08	−.02
A percent	49.73	46.63	48.25	48.68	16.17	15.65	.05	−.07
Variety	6.53	6.97	6.28	6.51	2.57	2.47	.05	.11
P	3.74	3.81	3.75	4.16	1.67	1.65	.00	−.12

[1] Group I: N_t=290; p_g=0.85. In class 44C.
[2] Group II: N_t=191; p_g=0.80. In class 44C.

(6) *Validity of weighted composite scores.*—An attempt was made to formulate a more readily employable procedure of evaluating Rorschach records. Certain categories were combined so that a composite score could be obtained. Each category was weighted on the basis of what it

would contribute to the classification battery and also on the basis of its variability. The final formula, expressed in standard Rorschach symbols, is:

Composite score = 2 (Dd+S percent) + 6 (FM) + 8 (W) − 1.5 (D percent) + 1 (R) − 1 (8, 9, 10 percent).

Data for a new group of pilots in primary training ($N_t=156$, $p_t=0.79$; in classes 44D and 44E) against a graduation-elimination criterion gave a biserial coefficient of 0.04. From this, it is concluded that the Rorschach scores used in this manner show no promise as a predictive instrument.

Evaluation

Several important qualifications arising from the nature of the experimental design limit conclusions concerning the validity of the Rorschach technique for classification purposes.

It is to be remembered that the present study did not validate directly the interpretive schema of the Rorschach test. Only clinical predictions, which were assumed largely to reflect these interpretations, were validated. In this technique, as in other clinical procedures, the clinical predictions were in part dependent on the pilot stereotypes of the examiners; only those qualitative features of the records which appeared relevant to their assumptions regarding pilot training were considered. In other words, at least to some extent, each examiner tended to employ his subjectively-derived system of weighting significant features of the records. It is not to be assumed that uniformity existed among the examiners with respect to their biases and weighting of factors. Differences between examiners, significant at the 1 percent level, have been demonstrated even for such objective data as number of responses obtained from examinees.

Since the test was originally intended for use with case-history data and other information, the present study cannot be considered to have defined the effectiveness of the technique when used in connection with a lengthier clinical procedure. It can be said that the objective category scores have been validated against the orthodox training criterion for pilots. It is hard to see how subjective evaluations based upon the same scores could yield much better results against the same criterion.

It is to be emphasized that the nature of the test itself, in its present form, precludes any final answers concerning its validity for predicting flying performance. As long as (1) the most important datum (qualitative interpretation) yielded by this test depends, to a critical degree, on individual insight, intuition, and skill, (2) the differences among examiners in these respects remain difficult to measure and control, and (3) examiner skill remains difficult to communicate, negative results may always be attributed to inadequacies in the examiner personnel. Such being the case, the test, in its present form, cannot be of practical use to a large-scale classification program.

There remains a possibility that the basic quantitative data yielded by the test can be subjected to another manner of analysis which would reduce or eliminate the influence of the examiner-difference variable and would validate significantly. Yet, it has been seen that one such type of analysis (the validation of single categories or of their composite) did not yield promising results. Some form of pattern analysis, in which the interaction of the personality variables (as represented by the scoring categories) is recognized and preserved, may very possibly prove to be promising. However, statistical procedures have thus far failed to provide a measure which would, in the first place, encompass the most significant features of the qualitative interpretations, namely, the interrelationship and interaction of personality factors, and in the second place, be capable of direct validation against a training criterion.

Group Administration of the Rorschach Test, CE701B

In an attempt to overcome the difficulties inherent in the time-consuming practice of individual administration of the Rorschach test, experimentation with forms suitable for group administration was attempted. Experimentation was undertaken first at Psychological Research Unit No. 3, and subsequently at Psychological Research Unit No. 1.

Description.—Two forms of the Rorschach test suitable for group administration were tried. The Picture Exercises test was developed experimentally at Psychological Research Unit No. 3.[4] No code number was assigned to this form. The Visualization Multiple Choice test, CE701B (Harrower-Erickson) (3), was validated at Psychological Research Unit No. 1.

(1) *Internal characteristics.*—The explanation of the two forms will be developed in parallel fashion. The basic apparatus used in the administration of the two forms was the same; a projector, a screen, and lantern-slide representations of the standardized Rorschach blots.

The main difference between the two forms was in the nature of the responses. In the Picture Exercises test each examinee recorded his free responses for each slide. As many responses as the examinee made were recorded for each slide. The main reason for this choice of technique was that the word responses supplied by Harrower-Erickson (2) were not regarded as necessarily the most applicable to the special sample of young American males undergoing selection for air-crew training. By allowing free responses in the first administration, too, it was hoped to develop lists of suitable alternatives empirically for use in later modifications of the test.

In the Visualization Multiple Choice test, the examinee's task was to choose responses from among the 10 standard Harrower-Erickson alternatives presented with each slide. If two of the alternates seemed to ap-

[4] Chief contributors: Lt. J. Richard Harsh and Lt. Wm. Stevens.

ply, the examinee was permitted to indicate a second choice. The alternatives for slide No. 2 afford an example of the type of choices presented:

1. A bug somebody stepped on.
2. Nothing at all.
3. Two scottie dogs.
4. Little faces on the sides.
5. A bloody spinal solumn.
6. A white top.
7. A bursting bomb.
8. Two elephants.
9. Two clowns.
10. Black and red.

(2) *Administration.*—The Picture Exercises test was administered to approximately 140 men at a time. Each examinee was given a prepared answer sheet on which to record his free responses. After detailed instructions, the room was darkened for 30 seconds, and a Rorschach blot was projected on the screen. Then lights were turned on so that there was sufficient light by which to record responses, yet leaving the blot dimly visible on the screen. Total testing time was 30 minutes. In the instructions the examinees were told that there were no right or wrong answers; that they were merely to write down all that they saw in each slide. The time limits and manner of presentation of the slides (30 seconds for study; 60 seconds for recording responses) were explained. The responses were written in long hand; accordingly, machine scoring of this form was not possible.

The Visualization Multiple Choice test (Harrower-Erickson) was administered to approximately 80 aviation students at a time. This form employed an answer sheet suitable for machine scoring. In the instructions the examinee was informed of the general nature of the procedure to be followed and was instructed specifically in the use of the answer sheet.

First, you will take a good look at each picture as it is shown and see whether it, or any part of it, reminds you of anything or resembles something you have seen. Then you will read through a list of suggested replies to see which of these is the best description of the blot.

Thirty seconds were allowed for the study of each blot, and then 30 seconds for the recording of responses. During the period when responses were being recorded, the blot was still dimly visible on the screen. Total testing time, including administration, was approximately 15 minutes.

(3) *Scoring.*—Because the Picture Exercises test is a new technique for administering the Rorschach ink blots, an arbitrary list of categories was adopted to which responses were assigned. The list follows:

A. Human.
B. Human anatomy.
C. Man made objects.
    1. Food.
    2. Animal anatomy and man made objects.
    3. Maps.

D. Animals.
E. Animal anatomy and animal detail.
F. Mythological and cartoon.
G. Marine.
H. Nature.
I. Microscopic.
J. X-ray pictures.
K. In complex responses, only the first mentioned objects are classified.
L. Original. In this classification system, means any unclassified response.

In this form all responses were recorded on cards according to category. Statistical treatment was undertaken only with first responses. Popular responses were determined by frequency counts.

The Visualization Multiple Choice test was scored by six trained Rorschach examiners, using the Harrower-Erickson method of scoring (2). The number of normal responses to each card, as defined by Klopfer and Kelley (4), is the principal category developed for statistical treatment.

Statistical results.—Because of the involved nature of the scoring process, only limited data are available.

(1) *Test reliability.*—No reliability data are available for the Picture Exercises test. Data are available, however, for the Visualization Multiple Choice test. The product-moment correlation of normal responses between the five odd-numbered cards and the total 10 was found to be 0.85. Correcting for overlapping,[5] the odd-even reliability was estimated to be 0.42. From these results it is obvious that the test does not have acceptable odd-even reliability. This is due to the small number of cards and, probably, also to the fact that the cards in the test do not appear to be of equal value for eliciting normal *v.* abnormal responses.

(2) *Test validity.*—The Picture Exercises test was validated on a sample of 591 pilots in primary training, in class 44D. The results for the six most predictive categories, as defined by Klopfer and Kelley (4), are presented in table 24.6.

The Visualization Multiple Choice test was validated against the criterion of graduation-elimination for 811 pilots in primary training. Validities determined on the basis of first and second choice, normal and abnormal responses, and of total number of second choices made were uniformly low (−0.14 to 0.06).

Evaluation.—On the basis of the data obtained for group administration of the Rorschach test, it is apparent that the multiple-choice form (Harrower-Erickson) would not contribute significantly to the prediction of graduation or elimination from primary pilot training. In the free-response form, the popular responses gave some promise, but this result needs verification with a larger sample.

[5] By means of the formula for correlation of a part with the remainder of a whole. The formula employed was:

$$r_{oe} = \frac{r_{eT}\sigma_T - \sigma_e}{\sqrt{\sigma_T^2 + \sigma_e^2 - 2r_{eT}\sigma_e\sigma_T}}$$

in which *o* = odds score, *e* = evens score, and *T* = total score.

Table 24.6.—*Validation of the six most predictive categories of the Picture Exercises test, using a graduation-elimination criterion for a sample of pilots in primary training*[1]

Category	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]
Popular responses	4.21	3.62	1.37	[3]0.21	0.24
Percent animal responses	45.90	41.60	15.60	[4].14	.11
Rejections	.65	.96	1.06	[4]−.14	−.15
Total number responses	13.56	13.98	5.12	−.04	−.01
Movement responses	2.25	2.24	1.37	.00	.02
Human responses	1.84	1.96	1.22	−.05	−.04

[1] $N_t = 591$, $p_g = 0.92$.
[2] Assuming an unrestricted stanine standard deviation of 2.00.
[3] Significant at the 1 percent level.
[4] Significant at the 5 percent level.

It is interesting to note that most of the examinees were not aware of the purpose of either test. At the conclusion of testing, one group of examinees was asked to volunteer opinions as to the purpose of the test. Some of the responses were: "A test for detecting camouflage; a test to determine visual acuity; a color-blindness test; a map-reading test; a test of imagination." Discussion among some of the examinees revealed that the majority tended to choose only the most acceptable social responses, as fear of consequences (elimination from training) in the Army situation made a completely free choice almost impossible. Reference to sex organs as one alternative provoked considerable hilarity, but even though this alternative might have been readily discerned in the ink blot, the majority of examinees refrained from making such a choice.

The Thematic Apperception Test, CE706A

It was hypothesized that the Thematic Apperception test promised types of information regarding the relationship of goals, values, conflicts and anxieties, and emotional complexes to air-crew success beyond that obtainable by the inventory procedures. Accordingly, studies were undertaken with a form of the test that is suitable for group administration.[5]

Description.—The present test is an adaptation of the Thematic Apperception test, developed at the Harvard Psychological Clinic (5). It is a technique that is purported to reveal to the trained interpreter indices of the dominant drives, emotions, attitudes, and behavior patterns of a personality. In clinical practice it is claimed to have demonstrated particular value by its power to uncover underlying tendencies which the examinee is either unwilling to expose or unable to expose because he is unconscious of them. The procedure of the test is to present to the examinee a series of pictures, each portraying one or more human beings who can be variously interpreted as to their characters and situations.

[5] Developed at Psychological Research Unit No 1. Chief contributors to the amended scoring technique: Lt. John S. Harding, Cpl. Charles E. Orbach, and Staff/Sgt. John L. Wallen.

The examinee's task is to tell a story about each picture in which he should answer such general questions as:

> What has happened to the individuals in the picture?
> What are their present thoughts and feelings?
> What will be the outcome of the story?

The second phase of the test consists of an interview in which the examinee is probed for his associations and memories in connection with various elements in his stories. Specifically, the test furnishes data concerning (1) the type of individual adjustment to different areas of life situations (for example, family, heterosexual, and authority relations), and (2) attitudes, motives, and emotions accompanying these adjustments. Thus, the test samples both covert and overt areas of personality.

(1) *Internal characteristics.*—For group testing, the pictures were transferred to slides and projected on a screen. The 12 slides consisted of the Harvard Psychological Clinic Thematic Apperception test pictures (5) of which the following examples are typical:

(*a*) Several male figures in work clothes, reclining on grass, and apparently dozing.

(*b*) A farm scene, adult male and adult female against background of tilled fields. A girl with books in her arms in foreground.

(*c*) An elderly woman with back turned to a young man holding a hat in his hands.

(2) *Administration.*—This test was administered to 28 men in a group, with administration time of approximately 75 minutes. The instructions read to the group are as follows:

> This is a test of creative imagination. You are going to compose stories, but do not bother about how well your stories are written; you will not be graded on spelling, phrasing, or style. Literary ability is unimportant. This is not an intelligence or a judgment test. Your stories will not be scored as right or wrong. This test is merely to find out how good your imagination is when you are pressed for time. Here is what you are to do. You will be shown some pictures. You are to make up a story to go with each picture, that is, the ideas for your story should come from the picture. You can make up anything you please; you are to use your imagination freely, but you must tell a complete story from beginning to end for each picture. As you will have only 6 minutes to write each story, be careful that you do not spend your time merely describing the picture. Be sure to give a complete plot.
>
> Remember that you will have only 6 minutes for writing each story. Plan to use all your time in order to make your stories sufficiently detailed. To help plan your time adequately, you will be told when there are 3 minutes left, and again when only 1 minute remains.

Under the group administration of the test it was not possible to conduct the usual interview to elicit associations and memories in connection with the stories. Instead, the slides were shown again for 5 seconds each, and the examinees were asked to indicate their reaction to each picture in terms of a 3-point scale of pleasantness. These reactions were recorded on a standard IBM answer sheet. Column A was marked for a pleasant reaction, B for indifferent, and C for unpleasant.

(3) *Scoring.*—As planned, the story analysis would involve the following steps:

(*a*) Scoring each story for each of 38 traits as listed by Murray (5).

(*b*) Making notes on the qualitative elements of each story.

(*c*) Combining the results of (*a*) and (*b*) for the entire series of pictures in order to define the major tendencies of the examinee's personality.

(*d*) Making an estimate, on the basis of these results, of the examinee's chances for success, on the nine-point scale, in elementary pilot training.

Experience proved that the method of scoring using 38 categories was too expensive and unwieldy. It seemed that the factors were too refined and subtle to be assessed from only 12 written stories. Accordingly, a revised scoring system was prepared, using only the 20 traits from Murray's list that could be isolated more easily in the data.

Six interpreters dealt with the material. The noncommissioned officer in charge had been trained in this test in the Harvard Psychological Clinic. An attempt to obtain uniformity of interpretation of the records was made by having introductory training sessions for the interpreters, plus supervision of scoring by the noncommissioned officer in charge.

When the scored and qualitative data were completed, the interpreter wrote a report synthesizing the examinee's personality tendencies. With these tendencies formulated, the interpreter next considered their probable composite contribution to success or failure in primary training in terms of a prediction on a nine-point scale. To facilitate the formation of such a composite judgment from so many variables, the interpreter, upon completing the scoring, classified each of the 20 personality traits of the examinee among four categories (strong, normal, mixed, weak) on the criterion of the probable contribution of each trait to success or failure in elementary training.

To the extent that the traits tended to cluster in the normal and strong categories, a higher prediction of success was given, and to the extent that the traits tended to cluster in the weak and mixed categories, a lower prediction was made. It is important to stress that the clinical prediction was not based upon the frequency results alone, but included nonscorable qualitative aspects of the examinee's stories. Hence, the final predictive estimate was in no sense derived from calculations approximating a crude formula. Rather, the rating was based on a broad clinical-type judgment in which were synthesized all relevant qualitative aspects, including those which could be systematically stated in terms of frequencies as well as those which could not.

The 20 traits employed in the later analyses are:

A. Ego image:
   1. Sex identification.
   2. Age identification.
   3. Action initiative.
   4. Adequacy.
   5. Endings.
   6. Goal orientation.
   7. Super ego.

B. Emotional pattern:
   1. Emotional strength and control.
   2. Frustration tolerance.
   3. Aggression.
   4. Anxieties.
   5. Emotional maturity.
   6. Picture tone.
   7. Examinee's test orientation.

C. Social adjustment:
   1. Orientation type, in terms of autonomy.
   2. To father and authority figures.
   3. To mother figures.
   4. To young adult females.
   5. To young adult males.
   6. To the Army.

Statistical results.—Data are treated in two groups. The first follows the system of scoring according to 38 categories, and the second follows the revised scoring according to 20 categories.

(1) *Test reliability.*—Reliability data are not available. The method of obtaining such reliabilities probably should be by means of a test-retest or alternate-forms procedure.

(2) *Test validity.*—The validity of the predictions of success (graduation-elimination) in primary pilot training, based on interpretation of the data using 38 factors, is indicated by a biserial r of -0.05 for a group of 293 pilots ($p_e=0.89$). This coefficient is not significantly different from zero.[7]

The correlations between the predictions based on 38 traits and the predictions for each of three other clinical techniques are given in table 24.7.

TABLE 24.7.—*Correlations of clinical predictions based on three other clinical techniques with the predictions for the Thematic Apperception test based on 38 Traits*

Technique	N	r
Rorschach Interpretation (CE701A)	320	0.08
Interview (CE707A)	323	.12
Observational Stress Test (CE710A)	316	.11

[7] The source from which these data were obtained failed to give information concerning M_e and M_r.

On the basis of these findings it is concluded that ratings based on 38 traits were unsuccessful in predicting performance of pilots in elementary training. Also these ratings are unrelated to ratings based upon other clinical techniques.

The validity of the predictions based on interpretation of the data using 20 traits was also determined. For a group of 191 pilots in primary training, using a graduation-elimination criterion (p_s=0.81), the biserial r was 0.05, which is not statistically significant.

The correlations between the clinical predictions based on the 20-trait interpretation of the Thematic Apperception test and upon other clinical techniques are presented in table 24.8.

Table 24.8.—*Correlations of clinical predictions based on several clinical techniques with the predictions for the Thematic Apperception test based on 20 Traits*

Technique	N	r
Observation during PMT Rest Period, CE709A	175	−0.06
Personal Audit, CE431A	190	−.02
Interaction Test, CE425A	185	.08
Observational Stress Test, CE710A	189	.17
Interview, CE707A	190	.19
Rorschach, CE701A	190	.03

Three of the 20 personality traits, deemed to be basic to the interpretations, were validated separately. First, the total number of favorable or plus variations in each trait was correlated with the criterion, then the number of unfavorable or minus variations in each trait was similarly correlated, and finally the difference between the number of pluses and minuses in each trait was correlated with the criterion. These results are presented in table 24.9.

Table 24.9.—*Validity data for ratings based upon three personality traits of the Thematic Apperception Test,*[1] *CE706A*

Factor	Plus	Minus	Difference
Action initiative	−0.05	−0.04	−0.03
Adequacy	.00	.04	−.02
Emotional strength and control	.03	−.11	.09

[1] N_t=186, p_s=0.80.

On the basis of these data it is concluded that ratings based on 20 traits likewise have no value in predicting success in primary pilot training. Neither do they correlate substantially with other predictions of success in pilot training derived from other personality evaluations. Three special indicators of personality likewise failed to show pilot validity.

Evaluation.—From these results it is clear that the group administration of the Thematic Apperception test as applied cannot be used to predict success in primary pilot training. It is possible that the Thematic Apperception test actually measures personality adequately, but pilot aptitude may strongly overshadow the importance of personality factors in elementary training. In other words, another criterion in which temperament plays a larger role, such as combat performance, might yield a higher validity. If temperament is a significant aspect of success or failure in pilot training, however, it would seem that anything as searching and global as the thematic test should predict the pass-fail variable.

Considerable difficulty is encountered in scoring this instrument. Because of the nature of the test, it is obvious that each record must be individually scored and interpreted. This interpretation required as much as 2 hours' time per case. Thus it is obvious that the test is not economical in its present form.

Probably one of the most important considerations, however, is the difficulty of interpretation itself. The bulk of the interpretational difficulties arose from three factors: (1) Examiner inexperience, (2) the subjective nature of the scoring, and (3) lack of secondary criteria. The nature of the first two difficulties is rather obvious. The difficulty arising from the lack of secondary criteria generally centered about the problem of discriminating as to whether an element was a projective or an introjective manifestation of personality. More specifically, were the stories the true projections of the examinee's personality, that is, drawn from his own personality, or no less likely, were the stories the wishful creations in the examinee's fantasy of what he would like to be and is not? In other words, the same type of hero in the stories of two different examinees admits of two different and opposite interpretations. Clearly, secondary criteria are needed.

The Rapid Projection Test, CE711C

It is taken as axiomatic that even the most normal person, if subjected to sufficient stress, will exhibit some personality disturbance. Some, however, are more susceptible than others to this. The term, combat neurosis, is generally taken to imply predisposing personality weaknesses, or proneness to break-down under stress. This projection test was constructed in an effort to obtain an estimate of susceptibility to combat neurosis.[a]

Description.—It was considered that some 22 characteristics seemed basic in forecasting combat neurosis:

1. Poor family adjustment.
2. Dependence.
3. Insecurity (uncontrolled).
4. Overcompensation.
5. Lack of group identification.
6. Inability to externalize hostile reactions.
7. Civilian functional somatic complaints.
8. Poor job adjustment.
9. Weak ego.
10. Schizoid tendencies (paranoid).

[a] Developed at Psychological Research Unit No. 1. Chief contributors: Pvt. Kenneth A. Fisher, Sgt. Harold M. Proshansky, and Capt. Donald E. Super.

11. Obsessive-compulsive tendencies.
12. Inability to take-it physically.
13. Lack of belief in democracy.
14. Lack of belief in support on the home front.
15. Lack of belief in worthwhileness of efforts.
16. Lack of belief in leaders.
17. Lack of advance awareness of combat.
18. Lack of interest in flying.
19. Uncontrolled prestige drive.
20. Inability to become absorbed in a technical job.
21. Inability to become detached from a situation.
22. Lack of conviction of personal invulnerability.

It was thought that this instrument might elicit information on some of these reaction patterns. Of these patterns, it was hoped that some would be critical indices of personality instability.

Two forms of the test preceded this one, gradually converging from a general exploratory treatment of the technique to a specific multiple-choice-type test in CE711C. In CE711A a relatively large number of pictures were taken from magazines or were sketched and made into lantern slides for group administration. The examinees were asked to write what they thought of each picture. This was intended to obtain protocols for each picture, so as to indicate the best multiple-choice items for each picture and to eliminate poor pictures. After a short period of experimentation, this form, having shown no indication of any positive results, was abandoned.

Form CE711B also served for the collection of preliminary data. It consists of 43 pictures with 3 questions concerning the content of each picture. The series of pictures is shown in sequence 3 times, with a 5-seconds exposure of each picture and 1 minute in which to write no more than a single sentence answer to each question. The series is shown to permit answering of the first question for each picture, then the series is repeated to permit answering the second question, and again for the third. For example, one picture shows a soldier lying on a beach alone, close to the water. The three questions are (*a*) "What is this soldier's nationality?"; (*b*) "Where are this soldier's companions?"; and (*c*) "Why is he lying on the ground?" No statistical treatment was accorded these results.

Form CE711C is derived from a Rapid Projection Test developed at the Harvard Psychological Clinic as part of an experimental battery for the selection of combat officers (6).

(1) *Internal characteristics.*—The material consists of 20 of the original 24 Murray Rapid Projection Slides. Four pictures were eliminated on an a priori basis, because they were very similar to other pictures. The slides are made from photographs of single male figures or, in a few cases, of groups of men.

Two series of answers are used for each slide, and answers are recorded on the standard A–O (15-choice) type IBM answer sheet. The first series consists of five possible interpretations of the content of the slide, any one of which will answer the question "What happened?" These are arranged in groups of five for each slide, picture one being 1—5; two, 6—10; and three, 11 — 15, etc.

The second series consists of 15 names or descriptions of feelings or emotions, any one of which will be an answer to the question "How is he feeling?" These are listed with letters from A to O. The examinee answers both questions with a single black mark on the answer sheet.

(2) *Administration.*—The test is group administered. It is presented to the examinees as a judgment test. They are told that it measures one's ability to size up persons at a glance. It is implied that the pictures represent actual situations and that in each case one of the answers is correct. Each picture is projected in a darkened room for 6 seconds. Then 1 minute is allowed to record answers with room lights on and the picture only dimly visible.

The following is the series of questions for a picture showing an elderly man sitting on a bench which appears to be on a ship. His chin is resting on his hands.

What has happened?

11. He has lost his job.
12. His wife has died.
13. His son was lost while on Atlantic convoy duty.
14. A splendid opportunity has just been offered him.
15. He has just learned that he has an incurable disease.

Feeling choices:

A. Daring, cocky.
B. Scornful, contemptuous.
C. Depressed, sad.
D. Pleased with himself.
E. Anxious.
F. Terrified.
G. Pained.
H. Angry.
I. Submissive or resigned.
J. Frustrated, or blocked.
K. Confused, hesitant.
L. Overjoyed.
M. Tired.
N. Amused.
O. Humiliated.

Statistical results.—Only item-validity data are available for this test.

(1) *Item validation.*—Cross-validation data were secured for sample of 360 graduates and 196 eliminees from primary pilot training in classes 44C, 44D, and 44E. The sample was split into odds and evens groups. Because 75 response spaces are available for each picture, the number of individuals selecting any one response is too small to permit computing a reliable item statistic. Two separate item analyses were made, therefore, one treating only judgments of factual content, and the

other, only judgments of emotional content. The data are presented in tables 24.10 and 24.11.

TABLE 24.10.—*Frequency distributions of phi coefficients for the criterion of graduation-elimination from primary pilot training, for the Rapid Projection Test, CE711A, based upon judgments of emotional content*

ϕ	Odds sample[1]	Evens sample[1]	ϕ	Odds sample[1]	Evens sample[1]
0.28–0.32	0	1	−0.02–0.02	15	13
.23– .27	0	0	−.07– −.03	22	16
.18– .22	1	1	−.12– −.08	3	6
.13– .17	4	2	−.17– −.13	0	2
.08– .12	6	7	−.22– −.18	1	0
.03– .07	11	17	−.27– −.23	0	1

[1] $N_t=278$, $p_e=0.65$.

TABLE 24.11.—*Frequency distributions of phi coefficients for the criterion of graduation-elimination from primary pilot training, for the Rapid Projection Test, CE711A, based upon judgments of emotional content*

ϕ	f		ϕ	f	
	Odds sample	Evens sample		Odds sample	Evens sample
0.28 — 0.32	0	1	—0.07 — —0.03	14	12
.23 — .27	1	1	—.12 — —.08	5	8
.18 — .22	2	1	—.17 — —.13	4	0
.13 — .17	2	3	—.22 — —.18	2	2
.08 — .12	5	6	—.27 — —.23	0	1
.03 — .07	19	22	—.32 — —.28	0	1
—.02 — .02	20	17			

In interpreting these data, it should be noted that for an N of 278 a phi coefficient of 0.12 is significant at the 5 percent level of confidence, and a phi of 0.16 at the 1 percent level of confidence. In the item analysis based upon judgments of factual content, for the evens group, only eight responses were significant at or beyond the 5 percent level. For the odds group, there were seven such responses.

For the item analyses based upon judgments of emotional content, there were 11 and 15 responses for the evens and odds groups respectively, that yielded phi coefficients significant at or beyond the 5 percent level.

The responses showing significance for one subsample did not show significance in the other subsample.

Evaluation.—It must be kept in mind that this instrument was designed to predict combat neurosis, and not merely for air-crew selection purposes. Whether the predictions were to have stemmed from a total score or from an over-all impression is not clear.

In either case, it is anticipated that this test would not have a high validity for predicting combat neurosis. The difficulty with this form of the test, according to Murray, is that no satisfactory multiple-choice answers have been developed to measure the variables thought to be essential for predicting combat neurosis.

The item validation results indicate that this test has no validity for the primary pilot training criterion.

703320—47—42

Empathetic Response Test, CE715A *

The rationale put forward by the constructor of this test is that a satisfactory measure of the examinee's predisposition to combat neurosis may be obtained from the manner in which he "empathizes" into a fictional character's affective state. The term empathy is used here to mean the imaginative projection of one's self into the mental state of another person.

Description.—The examinee reads a short story, written in such a way that he must empathize into the affective state of the chief character in order to supply a conclusion for the story. Each story contains a number of inconclusive clues or leads indicating a variety of possible endings, but no specific ending is supplied. They are written in a subjective style; that is, each is written as the conversations, musings, or reminiscences of the chief character in the story. The stories deal with various aspects of military life which are familiar, interesting, and significant to the typical soldier, but are combined into novel puzzle situations. It was hoped that the tendency would be for the examinee to supply a conclusion to the story compatible with his previously established and predisposing, affective habit patterns.

After the reading of the stories, two classes of questions are asked: Class I questions immediately follow each story and concern only that story. They do not suggest any specific story ending, so as to avoid prejudicing the examinee. Class II questions are all listed at the end of the test booklet, and they inquire as to the specific endings given the stories by the examinee. An example story follows:

At last, the hospital! They'll do something soon to ease the pain. That morphine shot Joe gave me didn't last long. My foot * * * leg * * * they hurt like hell, but not much damage. My hand. A mess! With a bun and mustard I could have a hamburger. Joe shook his head when he looked at it. Does that mean * * * ? God, what would Mary think? Ah, hell! Joe is smart, but he's no doctor. Could he tell? That doc's coming now * * * what will he think? My foot and leg * * * just a glance. Good, I thought they were O.K. My hand! He's gentle as a woman with it. Does that mean it is bad? I hope that he's not the doc they call Hacksaw. He is calling that other doc over to look. Why? Now they're going away. They keep looking over here. Why don't they quit mumbling? One keeps shaking his head. Does that mean he does or doesn't want to? Now they seem to agree.

This operating table feels swell * * * but it must mean an operation! No use asking the doctor anything because he won't give a straight answer. Probably ether for me soon. Yeah, here it comes. Ether! Does that mean a serious operation? How serious? Why did that thing have to explode and get me in the hand? This ether * * * it stinks. My ears are buzzing. Head's swelling * * * throbbing * * * swelling * * * swell * * *

Head * * * foggy. Can't think straight. I'm out of the operating room, in a ward. Operation's over? Must be! Operation * * * oper * * * hand! What did they do to it? Can't see my hand. Must be under the blanket. My arm is. Is * * * is my hand? No feeling in it. Should there be? Could there be? I'll

* Developed at Psychological Research Unit No. 1. Chief Contributor: S/Sgt. Robert F. Dice.

pull it out and look. I'll look when I get guts enough • • • if I get guts enough. Guts enough? I'll look now. Right now. Now!

God!

Two sample class I questions follow:

The soldier feels that the operating doctor:

A. Did not listen to the reasonable advice of the second doctor.
B. Felt unable to make a decision alone.
C. Performed the type of operation obviously needed under the circumstances.
D. Was chiefly interested in getting the operation done and over.
E. Was unconcerned about the result of the operation.

Choose the word or phrase which, you feel, best describes the general attitude of the soldier before the operation.

A. Fatalistic.
B. Frantic with worry.
C. Completely unconcerned.
D. Deeply concerned.
E. Emotionally detached.

The class II question for this story is:

In the story of the soldier wounded in the hand, I feel the operation ended in:

A. Loss of the whole hand.
B. Loss of part of the hand.
C. Restoration of the hand to partial usefulness.
D. Restoration of the hand to full usefulness.
E. A result concerning which I cannot reach a conclusion.

The test consists of seven stories, each followed by five questions of the class I variety with five alternatives each. The class II questions, which were administered as a group, consist of one question for each story. Each question had five alternatives.

(1) *Administration.*—Forty minutes were allowed for the test, two forms of which were validated. Parts I and II are class I and II questions for the first form, while parts III and IV are similar sections of a comparable second form of the test. Pertinent administrative directions are:

The purpose of this test is to measure your ability to understand a complete situation when only part of the facts concerning it are known to you. It is not a test of your ability to analyze and to draw scientifically correct deductions. This test measures your ability to grasp quickly a complex situation from a minimum of given facts. Your knowledge of human nature and your ability to understand how and why people act as they do in perplexing situations is of great importance in this test.

• • • you will read a series of short episodes and answer questions concerning the people and events in the episode. Read each episode carefully but rapidly. When you have finished reading the episode, turn to the following page and answer the questions concerning the episode. Once you have turned to the page of questions do not turn back to the previous page • • •

(2) *Scoring.*—A total of eight scoring keys was constructed for eight different categories of responses. Since no suitable external criterion for the selection of items was available, the items were selected and the provisional keys designed solely on the basis of a priori assumptions. The provisional keys, together with the scoring formulas used for each,

are: (1) General anxiety (number of Rights); (2) repression (number of Rights); (3) belief in worthwhileness of effort (R−W+10); (4) fear of death or injury (R−W); (5) sex conflict (R−W+10); (6) social adjustment (R−W+10); (7) evasion-confusion (number of Rights); and (8) attitude toward authority (R−W+10).

Statistical results.—Data are available for two somewhat overlapping groups of pilots who took primary training.

(1) *Test validity.*—Validation data are available for two samples of pilots, using the graduation-elimination criterion from primary training. The results of one group are based on parts I and II of the test, while those of the second group are based on the comparable parts III and IV. These data are presented in table 24.12.

TABLE 24.12.—*Validation data for two samples of pilots who took primary training, using the graduation-elimination criterion, for the Empathetic Response Test, CE715A*

Parts	Score	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
I and II[2]	General anxiety	8.13	8.34	2.65	−0.05	−0.04
	Repression	3.14	3.23	1.92	−.03	−.01
	Worthwhileness of effort	10.95	10.99	1.63	−.02	−.04
	Fear of death or injury	3.27	3.14	1.23	.07	.07
	Sex conflict	12.48	12.77	2.17	−.08	−.10
	Social adjustment	9.24	9.54	2.05	−.09	−.09
	"Evasion-confusion"	5.26	5.19	4.22	.01	.02
	Attitude toward authority	10.85	11.07	2.27	−.06	−.07
III and IV[3]	General anxiety	10.50	10.71	2.73	−.05	−.07
	Repression	6.33	6.18	2.27	.04	.04
	Worthwhileness of effort	13.49	13.26	3.03	.05	.04
	Fear of death or injury	8.03	7.82	3.14	.04	.02
	Sex conflict	11.34	11.35	1.06	−.01	.00
	Social adjustment	12.72	12.71	2.54	.00	.02
	"Evasion-confusion"	2.78	2.69	2.32	.02	.01
	Attitude toward authority	14.32	14.31	1.90	.00	−.01

[1] Corrected to an unrestricted stanine standard deviation of 2.00.
[2] For this sample N_t=491, p_g=0.68.
[3] For this sample N_t=537, p_g=0.67.

For parts I and II a biserial coefficient of 0.12 is required for significance at the 5 percent level and of 0.15 for significance at the 1 percent level. For parts III and IV, coefficients of 0.11 and 0.14 are required at the corresponding levels. None of the coefficients in table 24.12 approaches significance.

(2) *Correlation between parts.*—The correlations between paired scores of parts I and II, and parts III and IV, for a sample of 443 pilots in primary training who completed all four parts of the test, are presented in table 24.13.

TABLE 24.13.—*Correlations between paired scores of parts I and II, and parts III and IV, of the Empathetic Response Test, CE715A, for 443 pilots who took primary training*

Score	r'_{12}	r_{12}	Score	r'_{12}	r_{12}
General anxiety	0.06	0.11	Sex conflict	.05	.10
Repression	.32	.48	Social adjustment	.14	.25
Worthwhileness of effort	.10	.18	Evasion-confusion	.50	.67
Fear of death or injury	.08	.15	Attitude toward authority	.05	.10

The results obtained are disappointing, as only two of the keys show a level of reliability at all worthy of consideration. The evasion-confusion key scores only the "cannot decide," the "do not know," "do not understand," or "something not listed above" type of answer. The second key, repression, which seems to have some reliability, contains many of the same types of "cannot decide" responses and probably measures much the same thing as the evasion-confusion key.

Evaluation.—Upon the basis of the obtained validities, neither parts I and II nor parts III and IV of the Empathetic Response test have any validity for predicting success in primary pilot training. It is possible, however, that this instrument would yield a positive measure of susceptibility to combat neuroses, more readily than the picture-presentation projective type of test, because the examinee should empathize into an interesting story, written in subjective style, more readily than into a picture of doubtful meaning and interest.

For purposes of administration, in its present form the test has several advantages: it is easy to administer and relatively easy to score; it has considerable face validity, because the stories and questions deal with situations involving psychological stress or conflict in a military context.

Additional Use of Projective Techniques

In this section will be described briefly four techniques that have achieved the status of tests without having been administered. Only a brief rationale and a description of the items will be presented.

Picture Evaluation Test, CE712A [10]

The purpose of this test is to sample certain attitudes that are believed to be indicative of personality factors predisposing to combat neurosis.

(1) *Description.*—This test is a modified projective technique attempting to eliminate individual differences in verbalization. It was constructed for group administration. Each item consists of 5 pictures—a large unstructured picture, which is presented by means of a lantern slide for 5 seconds, followed by 4 smaller pictures, each structured in ways such as persons with various personality weaknesses might interpret the large picture. The examinees are asked:

(1) Which small picture is most like the large picture?
(2) Which small picture tells the same story as the large picture?

The answers are set up for machine scoring and are recorded on the standard IBM answer sheet. Both questions are to be scored, but it is felt that question No. 2 will be the more fruitful.

As an example, one large picture shows an airplane swooping close over the head of a man in the foreground. The man's mouth is open as if barking an order, or screaming, etc. The background is indefinite.

[10] Developed at Psychological Research Unit No. 1. Chief Contributor: Capt. Horace B. Van Saun.

The small pictures for this large picture are:

(*a*) Man dressed as flier is approaching a plane (only tail assembly is visible) in a normal airport setting. (Normal or escape picture.)

(*b*) A German plane ground-strafing American troops. (Attempting to get an undue awareness of the threat of death.)

(*c*) An obviously civilian plane in a minor landing crack-up. (Attempting to tap lack of interest or fear of flying.)

(*d*) Bugler blowing bugle with a background of graves. (Attempting to reveal uncontrolled insecurity as expressed in interpreting any incident as a threat to the self.)

It was planned to construct at least 15 large pictures, each with 4 alternative small pictures, in the exploratory form of the test.

Picture Sequence Test, CE713A [11]

It is assumed that two factors are crucial in determining relative resistance to the stresses of sustained combat exposure: (1) Intensity of anxiety, generalized and concerning combat, and (2) extent of counteracting control over anxieties. This test was designed to measure predisposition to combat fatigue in terms of these two personality variables.

(1) *Description.*—This is a nonverbal test combining features of both the projective techniques and the multiple-choice word-association tests. Each item consists of a stimulus picture, which presents a situation containing potential elements of stress, followed by two sets of four pictures each. The task is to construct a story by selecting one picture from the first set and one from the second. The first set presents varying stress elements in a continuum of increasing severity. The second set presents different completions of the story which are assumed to uncover varying degrees of anxiety control. For example, one stimulus picture portrays a young man sleeping. The first set of pictures portrays: (1) dreams of being carried off by an eagle; (2) dreams of being chased by wild animals; (3) dreams of falling from a great height; and (4) dreams of being trampled under a horde of "GI" shoes. The second set of pictures are: (1) Young man sleeping with smile on face; (2) young man sleeping with terror on face; (3) young man sitting up in bed, tensed; and (4) young man sitting on edge of bed, face in hands.

It was planned to construct 40 large pictures, plus 30 diversionary dummies. The test is suitable for machine scoring.

A Structured Answer Projection Test, CE714A [12]

This test seeks to measure, by a rapid-projection technique, those attitudes which predispose airmen to combat neurosis or which tend to serve as antidotes. These include such variables as belief in immediate superiors, conviction of personal invulnerability, belief in worthwhileness of own contribution to the war effort, and poor social adjustment.

[11] Developed at Psychological Research Unit No. 1. Chief Contributor: Sgt. Leo Srole.
[12] Developed at Psychological Research Unit No. 1. Chief Contributor: Lt. Martin Singer.

Description.—Fifty pictures are presented to the examinees, who are required to select the best description of a picture from four multiple-choice responses. The four possible responses are selected to reveal projections ranging respectively from those most indicative of adjustment to those most indicative of maladjustment to training and combat situations. The preface to each picture and the composition of the picture itself attempt to make all examinees identify themselves with the same character and from their perspective as aviation students. The pictures are as follows:

(1) Ten pictures with obvious responses; the first two presented as samples and the other eight scattered throughout the tests to make the examinees feel they are taking a test of observational ability. These eight obvious pictures may also uncover the extreme personality deviates.

(2) Twenty original pictures ambiguous enough so that interpretation invites projection on the part of the examinees.

(3) Ten pictures, from the Rapid Projection test, CE711B, so that the results of the present technique may be compared to the results obtained with that instrument. Multiple choices are also used here.

(4) Ten pictures used in the Picture Evaluation Test, CE712A, for comparison purposes. Multiple-choice questions of the same order as described above are used.

As an example, the following questions are asked concerning a picture of an officer who has just pulled the rip cord of his parachute.

From his expression:

a. He is getting set for the shock of the parachute opening.
b. The parachute has opened and he is calmly floating down.
c. He is terrified, because the parachute has failed to open.
d. He is worried whether or not the parachute will open.

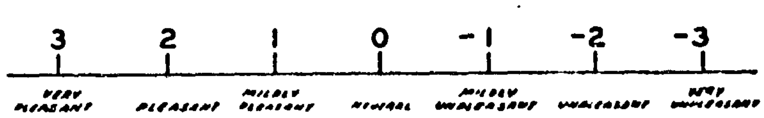


FIGURE 24.1
SCALE USED IN PICTURE JUDGMENT,
CE716A

Picture Judgment Test, CE716A [13]

It was hoped that this test would measure degree of fear and sensitivity to combat situations, and thus make a contribution to the prediction of susceptibility to combat neurosis.

[13] Developed at Psychological Research Unit No. 1. Chief contributors: Sgt. Harold M. Proshansky and Cpl. Walter J. Reis.

Description.—The examinees are asked to rate 150 pictures projected on a screen.

These pictures will show 24 pleasant nonmilitary situations, 25 neutral nonmilitary situations, 25 ambiguous nonmilitary situations, 38 unpleasant nonmilitary situations, and 38 combat situations.

It will be established statistically that the pictures finally selected have, for the majority of aviation students, the qualities: Pleasant, neutral, ambiguous, and unpleasant. It is expected that, canceling out the constant factor of strong or weak judgments, aviation students who are more sensitive to combat situations will (1) either mark a greater number of combat pictures in the unpleasant categories than their comrades or (2) avoid the situations by marking combat pictures neutral.

OBSERVATIONAL TECHNIQUES

Observational techniques yield data on the manner in which the examinee performs a task and provide estimates of his attitudes. In general, the primary function of observational data is to furnish descriptions of the examinee's overt behavior and interpretation of that behavior. The observer is considered to be a sort of complex measuring instrument who integrates his interpretations of the various observed reactions of the examinee into a complex judgment, which is recorded. It was felt that the observational techniques would yield data regarding personality traits that would supplement similar data obtained by other clinical techniques. Their independent value in prediction of air-crew success was also to be determined.

Conference for the Interpretation of Test Scores and Occupational Background, CE707A [14]

The interview method was considered basic to any project employing clinical procedures. The purpose of the conference technique was to attempt to obtain and to analyze the candidate's immediate and remote experiences, so that classification-test scores could be interpreted in the light of his background, as well as to give an estimate of the candidate's chance of success in air-crew duties. The information to be obtained is not, as yet, available through any other technique.

Description.—The designation, "Conference for the Interpretation of Test Scores and Occupational Background," was given to this procedure in order to avoid the undesirable connotations of the term "interview," which might influence rapport with the aviation students. This interview technique permits the follow-up of leads and the asking of questions that are difficult to put on paper or, if placed on paper, are easily distorted by the examinee. The interview situation is made as informal and relaxed as possible, so as to elicit a maximum of information concerning the examinee.

[14] Developed at Psychological Research Unit No. 1. Chief contributors: Clinical Procedures Group.

(1) *Internal characteristics.*—The conference was conducted by one interviewer for one candidate at a time. It was guided by a standard set of questions and a list of the fields to be covered. During the interview, however, the manner of putting the questions was left to the discretion of the interviewer in order to encourage optimum rapport. Questions, for the most part, were general and gave the examinee an opportunity to enlarge upon any point. Few questions were aimed at obtaining purely factual information, because most of this type of data was obtained previously on the printed questionnaires (Occupational Classification Experience Blank, CE603A, and Aviation Experience Blank). Specifically, the interview was planned to reveal: (1) The relationship of the past history of the examinee to his present personality; (2) what he has learned in terms of his opportunity to learn; (3) the developmental sequences of his life history; (4) his occupational history in relation to his likelihood for air-crew success; (5) additional material to aid in the interpretation of scores and measures derived from the Thematic Apperception Test, CE706A, the Rorschach Test, CE701A, and other printed tests of emotion and temperament; and (6) estimates (on a rating scale) of the examinee's self confidence, to aid in the interpretation of scores on the confidence tests.

(2) *Administration.*—Each conference was of approximately 50 minutes' duration. The main emphasis in the content of the interview was in four areas: (1) Occupation, (2) motivation, (3) stability, and (4) adaptability. The conference was so administered that adequate information would be available for a written report to include information under the following categories:

(*a*) *Thumbnail sketch.*—Included characteristics of motivation, stability, and adaptability, as well as such factors as physical build, vigor, dress, coordination, language, response to interviewer, and excessive physical activity.

(*b*) *Occupational history.*—This section contained information supplemental to the Occupational Background Blank; job ambitions before the war; preference for type of job 5 years from now; age at which examinee first earned money; why a particular job was taken; and special skill and interests. Discrepancies between level of aspiration and level of achievement were noted.

(*c*) *Family relationships.*—This area contained a brief description of family composition; indication of areas in which examinee would prefer to rear children differently than in the manner in which he was brought up; type of work which his mother and father would prefer for him; and their attitude toward combat flying. Important in this section were critical points in familial adjustment; such as the possibility of destructive sibling rivalry, the effect of marital discord on the examinee's discipline, goals, and ideals.

(*d*) *Developmental history.*—This area was concerned with an evaluation of self in early childhood and in the preschool area; adequacy of "social adjustment"; and physical adequacy in terms of any remembered severe illnesses.

(*e*) *School experiences.*—Described the age at which school was begun and concluded, best and poorest subjects; likes and dislikes in the school situation; behavior or disciplinary problems; extra-curricular participation; level of aspiration in relation to school and the extent to which this was reached; competition with siblings; reactions to teachers; and ambitions in relation to further education.

(*f*) *Leisure time interests and hobbies.*—This section was a description of the organizations in which the examinee had been a member; how leisure time was spent; what things afforded the most pleasure; what things "got him down?"; participation in athletics and in community organizations, and level of participation (member or officer).

(*g*) *Socialization.*—A description was given of the most difficult problem faced; type of individual most admired; relation with girls; social skills possessed; role in the family; extent to which independence from family control has been obtained.

(*h*) *Army.*—This section described ambitions in the Air Corps; evaluation of preparedness for this ambition; estimation of chances of getting wings; feelings in regard to combat; reasons for applying to the Air Corps; plans if "washed out" of training; reactions of parents and friends to his being in the Air Corps.

(3) *Scoring.*—Three types of treatment were accorded the material obtained: (1) An interview report was written, following the several categories just enumerated; (2) predictions for success in air crew in general, and for the specialties of bombardier, navigator, and pilot were made, each on a nine-point scale; (3) ratings of confidence were made on a five-point scale: Complete lack of self confidence; underconfident; confident; overconfident; and complete overconfidence.

Statistical results.—Owing to the nature of the data, only a restricted number of statistical procedures were attempted.

(1) *Test reliability.*—Since any estimation of reliability based on this instrument due to the method of administration would involve several unwarranted assumptions, no reliabilities were computed. It would be possible to determine the reliability of ratings, if two or more raters were to make independent clinical predictions of success based on the information contained in the interview summary.

(2) *Test validity.*—Clinical predictions on a nine-point scale of success or failure in elementary pilot training are available for validation. These ratings represent an over-all evaluation of the interview material, made at the end of the conference and after the interviewer had completed his interview summary. The ratings of the interviewers were

converted, for purposes of comparison, into a common scale, with a mean of 4.5 and a standard deviation of 1.5. Validation results are given in table 24.14.

TABLE 24.14.—*Validation data for clinical predictions based on two groups of pilots in primary training using the graduation-elimination criterion, for the Conference for the Interpretation of Test Scores and Occupational Background, CE707A (an interview)*

N_t	p_s	M_s	M_e	SD_t	r_{bis}	${}_cr_{bis}$
293	0.89	...	...	...	0.06	...
190	.81	4.73	4.36	1.55	.13	0.14

Neither of these validity coefficients is significantly different from zero. Clinical predictions of air-crew success derived from the interview do not significantly differentiate graduates and eliminees from elementary flying training.

Ratings of self confidence on a five-point scale are also available. These ratings were an over-all evaluation of the examinee's confidence as manifested in his behavior during the interview and in the life-history data. The ratings were converted to a scale with a mean of 3.00 and a standard deviation of 0.75. The biserial validity coefficients of the confidence ratings were determined on a sample of 293 pilots in elementary training, 80 percent of whom graduated. The biserial coefficient, −0.07, was not significantly different from zero.

The correlations of the clinical predictions based on CE707A with those based on other techniques are generally low. These results for several samples of pilots in elementary training are presented in table 24.15.

TABLE 24.15.—*Correlations of clinical predictions based upon Conference for the Interpretation of Test Scores and Occupational Background, CE707A, with predictions based upon other techniques*

Technique	N	r
Interaction Test, CE425A	323 184	0.04 −.03
The Personal Audit, CE431A (Author's Predictions)	190	.18
Rorschach Test, CE701A-I (First Prediction)	189	.23
Rorschach Test, CE701A-II (Second Prediction)	320	.09
Thematic Apperception Test, CE706A	323 189	.12 .19
Observations During Psychomotor Testing Rest Period, CE 709A	303 174	.06 .13
Observational Stress Test, CE710A	316 188	.04 .01

Analysis of the factual material in the interview, or of the major areas of motivation, socialization, emotional stability, and occupational background were not attempted.

Evaluation.—The influence of examiner differences on the interview procedure and results was not determined in this study. There were several phases of the interview procedure in which the factors of skill and intelligence of the examiner might become a critical variable: (*a*) In writing the interpretive summary and evaluating the factual content of the interview for areas of personality such as *stability, maturity, adaptability,* and *motivation*; (*b*) in questioning and probing certain responses, taking up various leads to questions, as from suggestive remarks, and following them through; (*c*) in making the clinical predictions; evaluating the relevance and importance of various phases of the life-history data, relating them with the examiner's conception of the personality factors required for success in air-crew performance, and making a clinical rating on this basis.

The clinical predictions depended partly on an examiner's own views of the requirements of air-crew duties with respect to personality factors, and partly on his skill as an interpreter of the factual life-history material in relation to various areas of personality. In arriving at their predictions, the interviewers were forced to determine for themselves the relative weights to be assigned to the various facts in the life history. Facts of life history that were elicited also differed from one examiner to another and from one examinee to another. It seems important to note that the interpretations of the factual materials were in no way determined by uniform criteria or by a single theory of personality. It seems legitimate to question whether the validity data pertaining to the ratings reflect the true validity, either of the factual data of the interview alone or of the interpretive features of the interview relevant to various areas of personality.

The analysis presented here does not, by any means, exhaust the possible analyses or further uses of the data. The data were not used in connection with two of the original aims for which they had been intended, the final conference and case history.

Observation of Atypical Behavior During Psychomotor Testing, CE708A [18]

This technique is based on the assumption that both the psychomotor tests and the air-crew training situations involve the performance of complex tasks under stress. It is thought that evidence of unusual tension, confusion, or excessive emotionality observed in the testing situation might serve as a basis for predicting similar behavior in the air-crew training situation.

Description. (1) *Internal characteristics.*—This is not a test but rather a descriptive rating procedure utilized when atypical behavior is observed by the regular psychomotor examiner. The procedure requires the examiner to record any behavior that he considers atypical in each group of four examinees taking the psychomotor test simultaneously.

[18] Developed at Psychological Research Unit No. 1. Chief contributor: Maj. Glen L. Heathers.

It was believed that the experience of the examiner with the psychomotor test would provide a sound basis for deciding whether individual behavior indicated tension, confusion, or any other striking characteristic. The object of these observations was to isolate the individuals who were judged to represent the extremes of atypical behavior.

(2) *Administration.*—When atypical behavior was observed, the examiner wrote a brief description of it after the four examinees had left the testing room. The examiner also indicated the adequacy of his opportunities for observing the examinees.

In the use of this technique, two rules were observed: (1) The examinees were given no reason to believe that their conduct during the psychomotor tests was being observed; for example, the data sheets were concealed from the examinees, and (2) the observational data obtained by each examiner were independent from similar data obtained by other examiners; for example, the examiners did not discuss observational findings on any examinee with each other.

(3) *Scoring.*—The examiner rated only those men considered atypical. He checked appropriate categories on the rating sheet and also wrote a very brief description of the behavior observed after the appropriate categories provided on the data sheet.

Briefly the categories were: (1) Tension, the examiner indicating and describing undue neurotic tension; (2) confusion, including poor attention, misunderstanding of test task, or erratic performance; (3) verbalizations, exclamations, rationalizations, and nervous speech; (4) disobedience, including willful disregard of instructions; and (5) other. In addition, the examiner checked the adequacy of his observation as either good, fair, or poor.

Statistical results.—No statistical data are available on this procedure. Reliability coefficients were not determined. The basis for computation would be observation of the same examinee in a given test by two or more examiners. This was not done. Only occasionally did two or more examiners report atypical behavior on the same examinee. Validity data were not obtained. Preliminary analysis of the data indicated that atypical behavior was noted in approximately one out of eight examinees.

Observation During Psychomotor Testing Rest Period, CE709A [16]

The design of this procedure was based on the theory that a good measure of individual personality characteristics might be obtained in a relatively informal, social situation structured to evoke relatively spontaneous and uninhibited comments, expressions of attitudes, and behavior. It was hoped that the observational material gained in this manner would aid, eventually, in giving valuable material for use in case histories and in making composite predictions for each cadet based on all the clinical procedures.

[16] Developed at Psychological Research Unit No. 1. Chief Contributor: Maj. Glen L. Heathers.

Description.—CE709A is not a test in the sense of having standardized questions to be answered. Rather, the men are placed in a situation that is thought to be sufficiently provocative to cause spontaneous reactions to the psychological testing and to their fears and hopes in regard to flying.

(1) *Internal characteristics.*—The procedure utilizes a waiting station in the psychomotor test section for groups of four examinees during the 15-minute period immediately following the Finger Dexterity test. The group is observed, insofar as possible without their knowledge, by the examiner who participates in the social situation. In order to provoke conversation and reactions indicating significant motives or attitudes, certain objects calculated to cause comment were placed in the room. On the wall were a cadet recruiting poster and pictures from Life magazine representing the duties of bombardier, navigator, and pilot. In addition, there were scattered about the room a number of parts from wrecked airplanes—a broken propeller, a burned brake drum, and a number of broken and twisted instruments.

(2) *Administration.*—The examinees were taken to the waiting room for the 15-minute period following the finger dexterity test. The examiner was there ostensibly to supervise the filling out of appointment slips for the clinical group tests. In this way the examiner's function was not open to question. This procedure took no more than 2 minutes, which allowed time for the examiner to make the necessary observations and ratings. After the appointment slips were filled out, the examiner informed the men that they were to remain in the room for the remainder of the rest period but were free to sit or move around, to make themselves comfortable, and to talk as much as they liked. The examiners were cautioned (*a*) not to observe the students too cautiously, (*b*) to keep all data sheets out of the room, (*c*) to act as natural as possible so as not to arouse suspicion, and (*d*) to join in the discussions as much as necessary and ask informal questions such as, "How are the tests going?" in order to provoke discussion.

At the conclusion of the observation period, the four examinees moved on to the next psychomotor test. The examiner went to another room to write up his findings, while an alternate examiner made the observations on the next group.

(3) *Scoring.*—In the separate room the examiner filled out a check list of traits and then wrote a thumbnail sketch of each man. The check list included 22 traits: Participant, nonparticipant, leader, entertainer, withdrawn, stoiid, nervous, elated, depressed, curious, confident, profane, rationalizing, seeks group approval, seeks examiner's approval, worried, griped, tense, pleasing to examiner, annoying to examiner, idealistic, and realistic. Definitions of these categories were furnished so that ratings would be consistent. A double check was used to indicate an extreme degree of the trait.

The thumbnail sketch was aimed at defining the individual, and was intended to lead toward the rating of personality fitness for air crew. It was not to be a restatement of traits that were checked; it was intended to be more general and to be expressed in spontaneous terminology. The findings were to be used to support the rating of personality fitness for air-crew training which was made on a nine-point scale adopted in the Clinical Techniques Project.

Statistical results.—Observations were made on approximately 600 aviation students.

(1) *Test validity.*—The validation coefficients obtained for the clinical predictions are given in table 24.16.

TABLE 24.16.—*Validation data for clinical predictions based on Observation during Psychomotor Testing Rest Period, CE709A, using the graduation-elimination criterion*

N_t	p_s	M_s	M_e	SD_t	r_{bis}	$_cr_{bis}$
273	0.90	...	...	...	0.10	...
176	.82	4.23	3.66	1.51	.21	.22

Correlations with predictions based on other clinical techniques are shown in table 24.17.

TABLE 24.17.—*Correlations of predictions for samples of pilots in primary training based on Observations during Psychomotor Testing Rest Period, with predictions from other clinical techniques*

Name of technique	N	r
Interview (CE707A)	300	0.06
Observational Stress Test (CE710A)	293	.02
Interaction Test (CE425)	300	[1].16
Interview (CE707A)	174	.13
Rorschach 1st Impression (CE701A-1)	173	[2]−.17
Observational Stress Test (CE710A)	173	.06
Interaction Test (CE425B)	170	−.07

[1] Significant at the 1 percent level.
[2] Significant at the 5 percent level.

Evaluation.—On the basis of present statistical treatment, the results obtained with this observational technique indicate that it has rather questionable value for the prediction of air-crew success. Certain features of the test limited its usefulness: (*a*) The examiners making the observations and predictions had little previous experience or training with clinical ratings; (*b*) predictions were colored, in part, by examiner stereotypes concerning the personality requirements of air-crew duties; (*c*) the attitude of the examinees may have been conditioned by the number of psychomotor tests they had taken prior to the rating period; (*d*) since in many cases the rest period came after the students had taken a number of tests, the predictions in an undetermined number of cases were possibly based more on comments the students made about how well they had done than was true in other cases. Some of these deficiencies could obviously be corrected and others minimized. The clinical observations from this test were originally intended for use in a case-conference procedure, rather than being used alone, but due to the abandonment of that part of the project, no further clinical use of the data was attempted.

Observational Stress Technique, CE710A [17]

The observational Stress Technique was designed in an attempt to represent the type of stress the aviation trainee would experience in the instructor-student relationship during flying training. It was assumed that his performance under critical observation would be best measured by tasks requiring divided attention, selective responses, and relatively fine controlled manipulations. The hypothesis, then, is that the examinee's poise and self assurance or his tendency to become confused or blocked, as measured by this technique, is related to success in air-crew training.

Description.—Since this is primarily an apparatus test and will be described as such in another report of this series, only those aspects more pertinent to the stress technique as measured by ratings will be mentioned here.

(1) *Internal characteristics.*—The apparatus consists of seven examinee's controls mounted in a table. The examinee's task is to stop the hand of a clock by keeping all seven controls set correctly at the same time. The controls consist of a foot pedal operated by the right foot, a stick operated by the right hand, and five levers manipulated by the left hand. These five levers are an assembly of throttle, mixture, and propeller-pitch controls adapted from a light plane and an assembly of two controls set by a thumb catch. The correct setting of each of six of the controls is indicated to the examinee when a corresponding signal light is illuminated and of the seventh control when a corresponding buzzer shuts off. During the test, the correct setting for each control is changed frequently by the examiner, according to a standardized schedule.

(2) *Administration.*—One examinee at a time is tested by one examiner and one observer. The directions require the examiner to make standardized criticism of the examinee's performance in an attempt to increase the stress aspect of the situation.

The examiner and observer are separted from the examinee by a one-way-vision screen, so that the examinee may be observed without his being able to note any unstandardized reaction on the part of the examiner or observer. The examiner administers the directions and the stress part of the test. The observer records his detailed observations of the examinee's performance on a prepared data sheet, and also records 6 clock scores which give the error times on several of the controls during each of the three test periods.

[17] Developed at Psychological Research Unit No. 1. Chief contributors: Tech./Sgt. James C. Crumbaugh, Maj. Glen L. Heathers, and Lt. Frederick G. Tice.

When the examinee enters the room, he is told in a forceful and critical manner that during the test he will be under constant observation from behind a screen. Then he is told to seat himself so that he can manipulate the controls easily. The test sequence follows:

(*a*) Anticipation period (1 minute): Examinee sits for 1 minute without further instruction. If he touches the controls, speaks, or attempts to rise, he is told to remain seated at ease.

(*b*) Directions period (1 minute): Full instructions are given orally concerning manipulation of the controls. Since they have little significance in connection with the observational technique described here, the instructions will not be quoted.

(*c*) Test period A (3 minutes): During the first minute of this period, the examinee attempts to match only one pattern. During the next 2 minutes of this period, the pattern changes 4 times at standardized intervals of 30 to 45 seconds. At approximately the same time intervals, the examiner administers standardized stress directions:

Stop the clock quickly. Always set the R control first. Set the stick and pedal controls next and keep their lights on while you set the rest of the controls. Set the "T" control next after these three * * * If your movements are jerky you will get a very poor observation score. Keep that stick away from the side of the slot! * * * Turn the buzzer off * * * If you make the lights flicker, it shows us that you are tense * * * Watch yourself * * * A record is being made of every false move that you make * * * Be quick!

(*d*) Test period B (3 minutes): During the first 1 minute and 15 seconds the examinee has only one pattern to match, while during the last 1 minute and 45 seconds the patterns change 4 times in a standardized fashion as in Test Period A. Again the examiner administers standardized stress directions:

Set the controls! * * * You must work more quickly * * * Your scores are not nearly good enough yet. Remember we are rating you the same way a primary instructor would rate you on your flying * * * You will have to do things exactly right or you are through * * * Are you letting a simple test like this confuse you? * * * By now you should be able to get all the lights on quickly and get a good clock score. But how well are you really doing? Size yourself up honestly * * * You are still making too many errors.

(*e*) Rest period (1 minute): During this rest period the clock scores are recorded and observational ratings made. Stress directions continue:

* * * You will wait while we record your observation score * * * You are still under critical observation.

(*f*) Test period C (2 minutes): During this test period the administrative procedure is directed toward failure stimulation. The patterns are changed 6 times after intervals of 15, 20, 15, 15, 10, and 40 seconds, in order. Directions are:

Don't make lights flicker on and off. Be steady * * * quit making errors. You aren't moving fast enough * * * More speed * * * Hurry and stop the clock * * * Last chance * * - Set controls quickly * * * You are still making errors.

703320—47—43

(3) *Scoring.*—In this account, we are interested only in two types of data: Ratings of observed traits and a general prediction of air-crew success.

Ratings of behavior were made by the observer during each of the six periods of the test. During the test period the examinee was rated on such characteristics as general manner (poised, ill at ease, relaxed, tense, confident, confused, or blocked); comprehension of task (adequate, uncertain, or poor); operation of controls (erratic, smooth, hesitant, impetuous, cautious, exaggerated); reactions to criticism (obedient, ignores it, confused, annoyed, slow, or prompt); incidental behavior (frowning, absorbed, anxious, grinning). The rating consisted merely in noting whether or not the examinee exhibited these characteristics. In addition, whenever possible the observer wrote a thumbnail sketch to supplement and clarify the ratings.

Predictions of air-crew success were made by the observer and examiner together on a nine-point scale for probability of success in air crew, in general, and as a bombardier, navigator, and pilot, specifically.

Statistical results.—The results are based on a sample of classified student pilots.

(1) *Validation data.*—The observer's ratings of the subject's expressive behavior during the test periods were to be used in a final case conference. Since the case conference procedure was abandoned, these observational data were never utilized in a clinical manner.

The validity of predictions of success was determined only for pilot primary training. The data are given in table 24.18.

TABLE 24.18.—*Validation data for clinical predictions based on a group of pilots in primary training, using graduation-elimination criteria, for the Observational Stress Test, CE710A*

N_t	p_g	M_g	M_e	SD_t	r_{bis}	${}_cr_{bis}$
286	0.89	...	...	...	0.15	0.17
189	.81	4.59	3.75	1.47	[1].33	.35

[1] Significant at the 1 percent level.

The marked improvement in validity of the second over the first group of examinees is considered to be attributable to the observer becoming more efficient through practice in the prediction of eliminees. Since the clinical predictions are strongly related to the clock scores (r=0.61), it is possible that most of the validity of the ratings is due to the examiners' knowledge of how well the subject was doing rather than to observations of behavior apart from this knowledge.

(2) *Other data.*—The intercorrelations of clinical predictions based on this test with other techniques for the same examinees are given in previously presented tables in this chapter.

Evaluation.—From the data presented, three conclusions may be reached as to the utility of the Observation Stress Test, CE710A.

(*a*) Subjective predictions of success in elementary pilot training based on a clinical evaluation, as well as on other qualitative observations while the examinee is under critical observation in a stress situation, seem to possess significant validity; but these data are probably contaminated by indirect knowledge of test scores.

(*b*) As in other clinical procedures, the influence of the examiner-examinee-observer interaction variables on the validities of the clinical predictions are undetermined. All ratings were made by one examiner and one observer.

(*c*) The possible influence of objective clock scores in determining the clinical predictions has not been determined. If the strong correlation between clock scores and ratings represents a personality variable rather than coordination or skill, then it might be better to use the objective score (clock) to measure it in place of the ratings.

Control Confusion Test, CE214A [16]

This test embodies essentially the same characteristics as the Observational Stress Technique, CE710A, lacking only the verbal-stress directions which were administered in that instrument.

Description.—The same apparatus is employed in CE214A as in the Observational Stress Technique. It is administered by one examiner to one examinee at a time.

(1) *Scoring.*—In addition to recording the objective clock scores, ratings were made by the examiner. The examinee was rated on a three-point scale for comprehension of total task, smoothness of operation, flexibility, and tension. Speed of comprehension was rated on a 10-point scale. Prediction of pilot success was rated on a nine-point scale with the points defined as follows:

Points	*Probabilities*	*Verbal statement*
8	Almost 8 chances out of 8.	Success highly probable.
7	Almost 7 chances out of 8.	Success very likely.
6	Almost 6 chances out of 8.	A good bet for success.
5	Almost 5 chances out of 8.	A little better than 50–50 chance of success.
4	Almost 4 chances out of 8.	A 50–50 chance of success.
3	Almost 3 chances out of 8.	A little less than 50–50 chance of success.
2	Almost 2 chances out of 8.	A poor bet for success.
1	Almost 1 chance out of 8.	Failure very likely.
0	Almost 0 chance out of 8.	Failure almost certain.

Statistical results.—Observations were made on about 500 classified pilots.

(1) *Test reliability.*—Since each of the ratings was made only once, by one examiner, reliabilities could not be calculated. The basis for computation would be observation of the examinee in the test situation by

[16] Developed at Psychological Research Unit No. 1. Chief contributors: Maj. Glen L. Heathers and Tech./Sgt. James C. Crumbaugh.

two or more independent examiners, or by a retest procedure. Owing to the shortage of personnel, neither procedure was attempted.

(2) *Validity coefficient of ratings.*—Two samples of classified pilots in primary training were divided for purposes of validation. The division was made on the basis of odd and even calendar days of the month during which they had been tested. The results are shown in table 24.19.

TABLE 24.19.—*Validity coefficients of clinical ratings for pilots in primary school divided on the basis of odd and even calendar days of the month for the Control Confusion Test, CE214A, using the graduation-elimination criteria*

Rating	Day	N_t	p_g	M_g	M_e	SD_t	r_{bis}
1. Speed of comprehension	Odd days	228	0.85	6.5	5.8	2.5	0.15
	Even days	260	.88	6.3	5.0	2.5	.26
	All days	488	.86	6.4	5.4	2.5	.20
2. Comprehension of total task	Odd days	229	.84	2.9	2.5	1.3	.17
	Even days	263	.88	2.9	2.1	1.3	.34
	All days	492	.86	2.9	2.3	1.3	.25
3. Smoothness	Odd days	229	.84	2.3	2.1	1.2	.08
	Even days	263	.88	2.3	1.6	1.4	.24
	All days	492	.86	2.3	1.9	1.4	.16
4. Flexibility	Odd days	229	.84	2.5	2.3	1.2	.10
	Even days	263	.88	2.5	1.6	1.4	.33
	All days	492	.86	2.5	1.9	1.3	.23
5. Tension	Odd days	229	.84	2.4	2.4	1.1	−.01
	Even days	263	.88	2.6	2.0	1.2	.27
	All days	492	.86	2.5	2.2	1.1	.13
6. Prediction of Pilot Success	Odd days	230	.84	5.6	5.0	2.1	.15
	Even days	264	.88	5.5	4.2	2.3	.31
	All days	494	.86	5.6	4.6	2.2	.23

The intercorrelations of the six ratings are presented in table 24.20.

TABLE 24.20.—*Intercorrelations of clinical ratings for pilots in primary school based on the Control Confusion Test, CE214A*

Rating	I	II	III	IV	V	VI
I. Speed of comprehension	...	0.66	0.57	0.56	0.34	0.64
II. Comprehension of total task	0.66	...	.73	.70	.34	.19
III. Smoothness	.57	.73	...	.70	.43	.76
IV. Flexibility	.56	.70	.70	...	.37	.76
V. Tension	.34	.34	.43	.37	...	.55
VI. Prediction of pilot success	.64	.19	.76	.76	.55	....

Evaluation.—In a survey of the results obtained with the Control Confusion Test, CE214A, two points remain to be made in addition to the conclusions drawn from the observation-stress technique:

(*a*) There seemed to be an over-emphasis on skill in achieving the proper setting of delicately-balanced controls on the basis of visual stimuli. The relatively fine adaptations required of the examinee do not reveal very full information upon which to make clinical ratings.

(*b*) A plan, whereby more reliable ratings might be made through increasing the scope of the examiner's observations, would be to have a series of trials increasing in difficulty until a virtual break-down of performance is reached, after which a recovery trial would be administered. Clinical ratings based on such a performance might have more prognostic value.

The Interaction Test, CE425B [19]

This test was based on the assumption that semistandardized social situations are productive of significant behavior that can be rated and recorded by trained observers. German military psychology relied heavily on behavior in such situations as a source of data on leadership ability and other traits important in officers. Traits such as cooperation, resourcefulness, leadership, dominance, social adjustment, and self-confidence are those generally believed to be assessable in such situations. It was hypothesized, further, that these characteristics are related to success in air-crew performance.

Description.—The essentials of the semistandardized, social-situation test were: (1) A task to be performed; (2) persons with whom it was to be performed; and (3) an observer to record significant behavior. The test situation could have presented either a thought (verbal) problem or a concrete (overt) task. Exploratory work resulted in the conclusion that concrete tasks produce interactive movements within a group and encourage verbalization, while thought problems encourage silent mental processes with less directly observable behavior. This exploratory work was done as CE425A, from which the present test form emerged.

(1) *Internal characteristics.*—The task selected for solution was a modified Wiggly Blocks (7) problem. Four examinees were placed in the testing situation in which discussion and cooperation were essential for a satisfactory problem solution. The examiner recorded a chronological description of events and made ratings of the behavior of the examinees.

(2) *Administration.*—Three sets of noninterchangeable Wiggly Blocks were placed unassembled in a standardized fashion on a 2 x 3 foot table and covered by a cloth. The blocks were placed on a square table, shuffled, laid out flat, and parallel, with the ends pointing in their correct direction. Four examinees were seated, one at each side of the table. The examiner stood to one side while reading directions and at his starting signal removed a cloth, disclosing the test materials. The instructions required that the group plan the solution to the problem. No method was prescribed for solving it. Part of the instructions follow:

> * * * task which you are to perform as a group. This task consists of three separate puzzles. There are 27 pieces altogether, which you are to put together as quickly as possible to form three rectangular blocks of nine pieces each. You will have 10 minutes. Each of you will have two separate scores: One based on the time taken by the group to finish all three puzzles, and the other based on what you yourself do. You may solve the puzzles in any way you wish. You may talk during this test.

[19] Developed at Psychological Research Unit No. 1. Chief contributors: Cpl. Carl Glaser and Capt. Donald E. Super.

After 5 minutes elapsed, the examiner informed the group that they had 5 more minutes. Similar comments were made after 7 minutes ("3 minutes left") and after 9 minutes ("1 minute left").

Upon completion of directions, the examiner wrote a chronological history of the group performance and noted behavior of members of the group to be used as a basis for rating them.

(3) *Scoring.*—As in other clinical techniques, described in this chapter, predictions for success in air crew in general, and as pilot, bombardier, and navigator were made, each on a nine-point scale. These judgments were made by the examiner on the basis of (*a*) a survey of the check list of traits which was used to rate each subject, and (*b*) detailed descriptions of the student's test behavior.

The following traits were included in the check list and judged on a five-point scale: Cooperation, integration, dominance, aggression, emotional stability, and fertility of ideas. In addition to the ratings, qualitative notes were made in the form of a chronological description of events, in which was emphasized the role of each individual in his approach to the task, in its execution, and in his relations within the group.

The examinees seemed to fall into three general categories: (*a*) A small number assumed definite leadership, had fertile ideas, and worked effectively with the group, primarily as an integrating force; (*b*) a small number worked alone and were more destructive than helpful; and (*c*) the largest number apparently faced the task calmly and intelligently, were not outstanding influences in the group, but seemed to be helpful participants and occasionally leaders and integrators.

In the rated predictions for air crew in general, for bombardier, navigator, and pilot success, those of category (*a*) were given ratings of 6, 7, or 8; those of category (*b*), 4 or below; and (*c*), 4, 5, or 6.

Statistical results.—Observations were made on about 600 classified pilots.

(1) *Test reliability.*—Reliability coefficients were not determined.

(2) *Test validity.*—Validation coefficients obtained for the clinical predictions are given in table 24.21.

TABLE 24.21.—*Validation data for clinical predictions for groups of pilots in primary training, using a graduation-elimination criterion, based on the Interaction Test, CE425B*

N_t	p_e	M_p	M_e	r_{bis}	$_cr_{bis}$
293	0.89	....	....	0.17	0.19
186	.81	4.54	4.83	−.11	−.10

Neither biserial coefficient is significantly different from a zero correlation.

Validity of the check list of traits, upon which the clinical predictions were based, was made for about half of the group. These data are given in table 24.22.

TABLE 24.22.—*Validation data for check list of traits employed in the Interaction Test, CE425B, for a group of pilots in primary training, graduation-elimination criterion*

[$N_t=280$, $p_g=0.89$]

Trait	r_{bis}	Trait	r_{bis}
Cooperation	0.13	Aggression	−0.09
Integration	.07	Emotional-stability	−.17
Dominance	−.02	Fertility of ideas	.09

No analysis was made of the qualitative data in the chronological record of each group's performance.

Evaluation.—From the data presented, three conclusions may be reached as to the utility of the Interaction Test, CE425B:

(*a*) Check-list ratings of personality traits based upon observations of social behavior in a small group of men engaged in a semistructured group task are not valid for the prediction of success in elementary pilot training.

(*b*) Subjective predictions of success in primary pilot training based on an over-all clinical evaluation of the personality ratings, as well as other qualitative observations in a semistructured group task, are not valid.

(*c*) As in other clinical procedures, the influence of the examiner variable on the validities of the check list, as well as on the clinical predictions, is undetermined. It will be noted that in this test all ratings were made by one examiner.

The Relationship of the General Appearance of a Cadet to His Success in Primary Training

The purpose of this experiment was to validate appearance ratings of aviation students against success in primary training. The rationale for this experiment was not that physiognomy is directly related to success, but rather that prejudices might be operative in determining success or failure in primary training.[29]

It was believed that flying instructors, as well as officers conducting the adaptability-rating-for-military-aeronautics interview, have certain prejudices common to our culture, and that because of these prejudices fine-appearing aviation students might be more likely to graduate from flying training and to pass the ARMA than other aviation students.

Description.—The ratings in this study were made by the regular group test proctors during the course of administration of the written classification battery in the spring of 1943. Ratings were made only after the proctors had had several hours in which to observe the men whom they were to rate.

[29] Developed as an experiment at Psychological Research Unit No. 3. Chief contributors: Capt. Stuart W. Cook, Capt. Lloyd G. Humphreys, Capt. Robert Murphy, and Staff of Group Test Section.

(1) *Administration.*—Ratings were made on a three-step scale. Three independent ratings were made by three observers, for each examinee. Each rater had 60 examinees to rate during each session.

The proctors were instructed to make their ratings as if they were selecting new men for a fraternity. They were told to look at facial features, complexion, name, hair, bearing, and clothing, in arriving at their judgment.

(2) *Scoring.*—A rating of A was given an aviation student whose general appearance indicated that he was among the most likely to succeed as a pilot, a B was given the intermediate group, and a C was given those thought to be among the least likely to succeed. The letter ratings were assigned numerical values of 3 for A, 2 for B, and 1 for C, and then these scores for each student were summated to obtain a single value.

Statistical results.—Data are available for a sample of pilots who took primary training, and also for a sample who passed or failed the adaptability rating for military aeronautics.

(1) *Distribution of ratings.*—The observed distribution of ratings is contained in table 24.23. The expected distribution, assuming lack of correlation between raters, is also presented in this table.

TABLE 24.23.—*Observed and expected (assuming zero correlations between raters) frequency distributions of appearance ratings in the relationship of the general appearance of a cadet to his success in primary training*

Combined ratings	Observed frequencies	Expected frequencies	Combined ratings	Observed frequencies	Expected frequencies
9	93	32	5	570	681
8	315	242	4	262	221
7	709	710	3	115	29
6	844	995	N total	2,908	2,908

Chi-square for the discrepancy between these two distributions is 526.6, indicating a positive correlation between ratings.

(2) *Test validity.*—An estimate of the predictive value of the appearance ratings was made by validating the ratings against the criterion of success or failure in primary pilot training. Employing a sample of 2,228 pilots in primary training (p_s=0.82, M_f=6.15, M_s=6.15, and SD_t=1.34), the obtained coefficient between the ratings and the criterion was zero.

(3) *Relationship of ratings to ARMA scores.*—It was thought that appearance might be a factor in determining the ARMA score. The correlation between the two ratings is presented in table 24.24.

Evaluation.—The relation between ARMA scores and appearance ratings shows that the possible source of bias which is present in any face-to-face selection technique has been here relatively well-controlled. On the basis of this evidence, it is concluded that any validity found for face-to-face selection devices, that is, interviews or body-build measurement, may be free of the factor of personal appearance.

TABLE 24.24.—*Relationship of appearance ratings to the ARMA score*

Appearance ratings	ARMA		Total
	Pass	Fail	
9	89	4	93
8	308	7	315
7	681	28	709
6	791	53	844
5	542	28	570
4	238	24	262
3	108	7	115
N	2,757	151	2,908
M	6.08	5.71	
SD			1.33
r_{bis}			.13

It is clear that appearance ratings are unrelated to success in primary training, though very slightly related to adaptability ratings for military aeronautics. Though prejudices undoubtedly exist, the elimination procedure apparently minimizes such effects.

SUMMARY AND CONCLUSIONS

The original aims and assumptions underlying the clinical-techniques project and the clinical-type procedures were more comprehensive than the scope of the completed study. For various reasons, several major aims, fundamental to the thorough appraisal of the clinical approach, miscarried.

The present study can afford only a negative answer to the question of whether the clinical type of approach, in general, can be of value to the classification program. During the course of the study, emphasis shifted steadily from the attempts to develop a global type of analysis, based on a variety of clinical procedures, toward consideration of the specific clinical tests individually as predictive instruments and as means for getting leads for more objective tests.

In the techniques most characteristic of the clinical approach, summarizing predictions, based on a nine-point scale, of success or failure in primary pilot training, constituted the major datum for validation. The results show that the clinical ratings are consistently ineffective in the prediction of success or failure in primary flying training. This much gives us a categorical answer regarding clinical predictions of this type, based upon data such as were used in this project.

At least three sources of influence bear upon a clinical rating in any single instance: (1) Examiner opinions or stereotypes concerning the personality factors associated with success or failure in primary pilot training; (2) individual interpretations of the basic data (whether observational or projective in nature) with regard to various areas of personality; and (3) subjective (examiner) weighing of the personality characteristics thought relevant in estimating the chances for success of an aviation student. It should be added that little uniformity can be assumed for the various examiners with respect to their stereotypes, their interpretations, and the subjective weights they assign to personality factors. Large individual differences existed in the background, training, ability, and temperament of the examiners.

It should be emphasized that the attention given to examiner differences in this report underlines a basic difficulty in applying clinical procedures to a large-scale testing program. As long as (1) the most important datum of clinical procedures depends, to a critical degree, on individual insight, intuition, and skill, (2) the differences among examiners, in these respects, remain difficult to measure and control, and (3) examiner skill remains difficult to communicate or is "incommunicable," it will always be possible to attribute negative results to inadequacies in the examiners.

Another fundamental difficulty is that of validating directly the qualitative clinical interpretations. All attempts to express these interpretations in quantitative form inevitably exclude all, or certain, critical features of the interrelatedness and interaction of personality characteristics. It has been shown that clinical predictions of the type used in this study are unsatisfactory measures for this purpose in that they introduce both extraneous assumptions and examiner complications. The validation of single scoring categories (as in the Rorschach) is, at best, a superficial procedure, since it slights the basic dictum of the clinical approach which emphasizes global analysis and the interrelationship existing among the components of personality. It must be concluded that there has been little success with attempts to derive a measure of the most significant features of the qualitative interpretations (namely, the interrelationship, interaction, and balancing of personality factors) that would be capable of direct validation against the training criterion.

One of the original aims of the project was to develop a well-rounded clinical picture of the individual by means of a case conference in which a final prediction of success or failure in air-crew training would be based on the results of the entire clinical battery. A committee was to have made the predictions, using several sources of information all reinforcing or qualifying each other, which method possibly would have increased the reliability of the prediction itself as well as of the personality picture constructed from the clinical materials. The failure to carry through the case conference necessitated abandonment of another original aim, namely, to write case histories from which one could, upon receipt of validation data, build up a man analysis of pilot training. This might have afforded a total pattern for a group of scores, which could have served to negate a favorable individual score, compensate for an unfavorable score, or, in general, give additional meaning to any individual score.

One aim of the project was fulfilled, namely, the development of objectively-scored, group tests based on the principle of projection. The following tests related to this general purpose are: Rapid Projection

Test, CE711C; Picture Evaluation Test, CE712A; Picture Sequence Test, CE713A; Picture Judgment Test, CE716A; Restricted Word Association Test, CE702A; and the Empathetic Response Test, CE715A. In spite of their relative objectivity, none of these instruments validated significantly against a graduation-elimination criterion in primary pilot training.

The final conclusion, then, is that clinical predictions, summarized subjectively from the single clinical tests, are of little or no value for prediction of success in pilot training. A final answer, however, as to the useability of a thoroughgoing clinical approach for classification purposes must rest on solutions to the following problems: (1) The use of clinical procedures in combination, as in a case-conference technique, (2) pattern analyses which recognize and preserve the global approach and at the same time are capable of direct validation, (3) examiner variability, and (4) the possible use of other criteria, such as combat performance, which might disclose some value in the use of clinical techniques for selection purposes when training criteria do not.

BIBLIOGRAPHY

(1) Frank, L. K. Comments on the proposed standardization of the Rorschach method, *Rorschach Research Exchange*, 1939, 101–105.

(2) Harrower-Erickson, M. R. *Large Scale Rorschach Techniques*, Springfield: Charles C. Thomas, 1945, 419.

(3) Harrower-Erickson, M. R. Modification of the Rorschach method for use as a group test, *Rorschach Research Exchange*, 1941, 5, 145–153.

(4) Klopfer, B., and Kelley, D. G. *The Rorschach Technique*, New York: The World Book Co., 1942.

(5) Murray, H. A., *et al.* *Explorations in Personality*, New York: Oxford University Press, 1938.

(6) Murray, H. A. Note on selection of combat officers, *Psychosomatic Medicine*, Oct. 1943, 5, 386–391.

(7) O'Connor, J. *Born That Way*, Baltimore: Williams and Wilkins, 1928.

(8) Rorschach, H. *Psychodiagnostik*, 4th ed., Bern: Huber, 1941, 277.

CHAPTER TWENTY-FIVE

Measures of Specific Traits of Temperament[1]

INTRODUCTION

The two preceding chapters dealt with two approaches to measurement of temperament: personality inventories and clinical-type procedures. Both of these methods employ instruments designed to reveal information about a variety of aspects of human temperament. The primary objective of each test described in this chapter was to differentiate among individuals and distribute them on a single continuum. The tests discussed here, then, might be said to explore some of the component parts of temperament as contrasted with the more general approach of the preceding chapters.

It will no doubt be obvious to the reader that only a small proportion of the identifiable traits of temperament were explored. The decision to explore any given trait was made largely on one important basis. This was the apparent importance of the trait to air-crew success. Rationales for these decisions are covered in the discussions of the tests or groups of tests. In general, however, the presumption that a trait is related to air-crew success resulted from formal or informal job analysis. Job-analysis findings that suggested the areas covered by this chapter are cited in chapter 22.

MEASURES OF MASCULINITY

The tests discussed in this section, in common with certain other information tests, were designed to reveal some characteristics of temperament rather than characteristics of intellect of the individuals tested.

The Masculinity-Femininity Hypothesis

Studies of biographical-data and sports-and-hobbies items indicated that students who have acquired extensive knowledge of airplanes tend to be more successful in pilot training. Experience in riding motorcycles and handling guns also proved to be significantly correlated wtn pilot success. On the other hand, the same studies revealed that items concerned with art, literature, music, and the like, tended to yield negative validities for the pilot. Speculation concerning these results suggested the possibility that these differences could best be explained by the hypothesis that more masculine men tend to succeed in flying, while those inclined to femininity tend to fail. The possibility of determining a man's

[1] Written by T/Sgt. Paul C. Davis.

experience and interest in these subjects by means of information items instead of by biographical items, which are susceptible to falsification, was appealing. The success of sports-and-hobbies information items lent promise to this approach.

Additional evidence in support of the masculinity hypothesis was obtained in observation of the performance of bomber pilots in the European theater, which placed in bold relief the fact that the bomber pilot needs to be more than a good pilot. Among the characteristics apparently essential to the pilot's function as chief of the bomber crew are forcefulness, aggressiveness, and other attributes allegedly associated with masculinity. It appeared that those pilots producing the best combat results, those most successfully withstanding combat conditions, and those most successful in handling crews, all tended to exhibit the traits and habit patterns commonly attributed to masculinity.

Since insufficient evidence existed to establish masculinity-femininity positively as the determinant of validity in the studies cited above, it appeared advisable to test the hypothesis by the construction of a test or tests composed of distinctly masculine and feminine items or responses. The validity of classification of the items within this dichotomy would obviously influence the results of such a study. The judgments of psychologically trained personnel were accepted as the most practical, preliminary index to masculinity and femininity of items. Categories of information judged to be masculine or feminine by this method were listed and assigned to the various forms of the test, so that all items in a given category would appear in only one form. Analysis of results could then be studied by category if so desired.

The plan of research in the information area included the administration of the test or tests to senior students in high school for the purpose of determining the typical responses of the two sexes. In this manner responses could be designated as masculine or feminine on an empirical basis. Such an empirical key would make possible the desired validation of the hypothesis upon which construction of the test was based. Because of the tendency noted in earlier studies for the differentiation between masculinity and femininity in terms of information to approach zero as intelligence increases, it was planned to administer a general vocabulary test as a suppression instrument to use along with the information test. At the time this account was written, this part of the project had not yet been completed.

General Information (M-F), CE505GX4 [8]

As previously suggested, this test was designed as an expanded version of the valid portions in the biographical-data and the sports-and-hobbies inventories which appeared to differentiate on the basis of masculinity.

[8] The four forms, CE505GX4, 5, 6, and 7, were developed at Psychological Research Unit No. 3. Chief contributors: Sgt. David Grossman and other members of the test construction staff.

Description.—The information covered in this test is general, in the sense that it is commonly available. It is highly selected, however, in the sense that only information judged to be either predominantly masculine or predominantly feminine was used. Information judged to be equally common to both sexes was not employed.

(1) *Internal characteristics.*—Form CE505GX4 contains 100 items, 50 masculine and 50 feminine, as determined by subjective evaluation. Five categories of information judged to be masculine and four categories judged to be feminine are included. Table 25.1 lists the categories and numbers of items in each.

TABLE 25.1.—*Analysis of contents of General Information (M-F), CE505GX4*

Masculine items		Feminine items	
Category	Number of items	Category	Number of items
Comics	11	House and furniture	12
Hunting	9	Dancing	11
Trades	15	Cooking	16
Card games	9	Technical knowledge of music	11
Track (sport)	6		

(2) *Administration.*—The test is largely self-administering in view of its straight-forward informational character. No special directions or sample problems are employed.

(3) *Scoring.*—This test was first scored right (correct answers to masculine items) minus wrong (correct answers to feminine items) with all other responses scored zero.

Variations of the test.—Several forms of this test, differing in the categories used, were constructed.

(1) *General information (M-F), CE505GX5.*—Table 25.2 indicates the categorization of items in this form.

TABLE 25.2.—*Analysis of contents of General Information (M-F), CE505GX5*

Masculine items		Feminine items	
Category	Number of items	Category	Number of items
Explorers and inventors	12	Radio	12
Smoking and drinking	12	Literature	23
Horseback riding	13	French phrases	8
Animals and snakes	13	Cosmetics	7

(2) *General information (M-F), CE505GX6.*—Table 25.3 indicates categorization of items in this form.

TABLE 25.3.—*Analysis of contents of General Information (M-F), CE505GX6*

Masculine items		Feminine items	
Category	Number of items	Category	Number of items
Fishing	4	Movies	13
Photography	10	Art appreciation	13
Boxing and wrestling	3	Care of clothing	7
Household mechanics	5	Etiquette	2
Football	6		
Baseball	5		
Motorcycling	2		

(3) *General information (M-F), CE505GX7.*—Table 25.4 indicates categorization of items in this form.

TABLE 25.4.—*Analysis of contents of General Information (M-F), CE505GX7*

Masculine items		Feminine items	
Category	Number of items	Category	Number of items
General mechanics	10	Art appreciation	8
Horse racing	8	Music composers and titles	13
Water sports	7	Stenography	6
Boats	10	Clothes	14
Chess	5	Groceries	9
Farming and gardening	10		

Statistical results.—No data are available.

General Information, CE505GX3

This form is composed of items from exercise 3 of forms A and B of the Terman-Miles Attitude-Interest Analysis Test. Administration of the test to several thousand subjects of various age levels over a period of years resulted in the development by the authors of an empirical key differentiating typically masculine and typically feminine responses.

Description.—The information items contained in this test cover a wide range of subjects and interests, to which some of the responses are more common to males, some are more common to females, and a smaller number appear to be approximately equally common to both sexes. A few examples, in the form in which they were administered to Air-Forces personnel, are presented. Weightings of responses are indicated. The + indicates a masculine response, the − a feminine response, and 0 a neutral or ambiguous response. These symbols, of course, did not appear in the test booklet.

Peat is used for:

+A. Fuel.
−B. Pavement.
0 C. Plaster.
0 D. Road making.
−E. Don't know.

A buffet is used for:

0 A. Books.
+B. Clothes.
−C. Dishes.
−D. Food.
+E. Don't know.

'Pi' is equal to:

−A. 0.6666.
0 B. 0.7853.
+C. 1.453.
+D. 3.1416.
−E. Don't know.

(1) *Internal characteristics.*—In preparing the test for aviation students, some changes were made. Whereas the original test allowed for omission when the examinee believed he did not know the answer, this form provides a fifth, or E alternative, marked "Don't know." From the 140 items in the two forms of the Terman-Miles test, 75 items were selected as being most appropriate for use with aviation students.

(2) *Administration.*—The test is largely self-administering. The time allowed for answering the 75 items is 15 minutes.

(3) *Scoring.*—The authors' key for the selected items was adopted. This key includes a weight of +1 for each masculine response, −1 for each feminine response, and 0 for a neutral or ambiguous response, all based upon the empirical findings of the authors. The total score is the algebraic sum of the item response weights.

Statistical results.—Only item-analysis statistics are available, based upon unclassified aviation students tested in September 1944 at Psychological Research Unit No. 3.

(1) *Internal consistency.*—Based upon total score, the highest and lowest 200 (approximately 27 percent) of 750 papers were removed for item-analysis purposes. For this sample, 115 responses keyed as masculine yielded a mean internal-consistency phi of +0.11, a standard deviation of 0.09, and a range from −0.14 to +0.32. Likewise, 110 responses keyed as feminine yielded a mean internal-consistency phi of −0.10, a standard deviation of 0.12, and a range from −0.45 to +0.25.

Further light may be cast on the applicability of the key by examining the overlap of phi values between the positively and negatively scored responses. Of the 115 positively scored responses for which phi values were computed, 11 yielded negative phis with total score; and of the 110 negatively scored responses for which phi values were computed, 21 or about one-fifth yielded positive phi values.

Evaluation.—It is obvious from the data, therefore, that the authors' empirical keys could not be highly valid for the air-crew candidate population, since the internal consistency is relatively low. This is due in part, no doubt, to the fact that the original key was based on mixed samples, whereas the air-crew candidates were of one sex and age group only. A careful statistical study of responses of the air-crew candidates and those of samples of female population of similar age and background should make possible the development of a valid masculine key. Such a study should be followed by a validation of item responses against training and combat criteria.

Reaction Speed, CE451AX1

Since much earlier research in masculinity-femininity had been conducted by civilian psychologists, it seemed advisable to test the hypothesis by means of instruments developed by them. Terman, Miles, and

703320—47—44

Goodenough had quite thoroughly explored the word-association area and had developed tests reported to differentiate on the basis of masculinity-femininity.

The Terman-Miles and Goodenough word-association tests were administered to aviation students for validation. Because of the differences between the mixed civilian populations and the aviation students, it was obvious that empirical keys should eventually be developed for air crew. Pending validation of the items of the test, however, the authors' keys were employed for item-analysis purposes.

Description.—This is Goodenough's Speed of Association Test, a word-association test devised for the specific purpose of differentiating masculine-like from feminine-like individuals.

(1) *Internal characteristics.*—This test consists of 238 items of the Goodenough test plus 12 similar items added to make a total of 250. Each item consists of a stimulus word followed by a blank space. The space is provided for the response of the examinee. The stimulus words are all in common usage. Homographs are employed, since they offer greater latitude of interpretation for the examinee. Typical samples of items are given.

Fair ....

Park ....

Lead ....

Cast ....

Roll ....

Bear ....

Bust ....

(2) *Administration.*—Directions for this test are simple but important. Examinees are instructed to write in the space provided the first word or phrase suggested by a stimulus word, regardless of what it is. The fact that no right answers exist is also stressed. Instructions demand that a response be given to every item. The time allowed for the test is 20 minutes, plus 2 minutes for administration, Examinees are not allowed to go back over the items, even though they are finished before time is called. The time limit was so set that practically nobody finished.

(3) *Scoring.*—In view of the fact that free responses to the stimulus words are secured in this test, the scoring is difficult and time-consuming. The key covers approximately 150 pages and includes just about all possible responses to the stimulus words, listed either individually or by class.[8] Each possible response or class of responses is assigned a weight on an 11-point scale from 5F to 5M; with the midpoint designated as A (ambiguous). Male and female examinees receive different scores for the same response in many cases. The following is a portion of the

[8] We are indebted to Dr. Florence Goodenough, who kindly made scoring keys available.

key for marking responses to the word STAKE, showing the weights (1, 2, 3, etc.) for the sexes assigned for each response:

M	F	
2M	3M	All references to materials as wood, iron, etc., also stick of wood.
A	A	Synonyms as stick, post, pole, peg, stick in ground, etc.
A	A	All references as meat, food, or eating (confusion with steak).
2F	2F	Ground.
4F	2F	All references to games as horseshoes, croquet, play, etc.

Final score on the test is the total masculine less the total feminine score, making high positive scores masculine.

Statistical results.—No statistical data are available for this test.

Reaction Speed, CE451AX2 (Terman-Miles)

Terman and Miles utilized both free-association and multiple-choice association techniques and reported results of the two methods to be similar.

Description.—This is the original Terman-Miles word-association test with three stimulus-words deleted, on the ground that they might be objectionable to the aviation-student population.

(1) *Internal characteristics.*—The test contains 117 words, each of which is followed by four other related words, lettered A through D.

(2) *Administration.*—Directions for answering are very simple. A sample item is presented, and the examinees are instructed to look at all the alternative words and then select the one which seems to go most naturally with the stimulus-word. They are cautioned to answer quickly and not think too long about any one word. The samples given are typical of those appearing in the test.

FLESH	A. blood.	B. color.	C. meat.	D. soft.
DEVIL	A. dare.	B. evil.	C. hell.	D. tempt.
NEEDLE	A. compass.	B. eye.	C. pine.	D. sew.
HUNT	A. find.	B. gun.	C. search.	D. shop.

(3) *Scoring.*—The total score on the test is the algebraic sum of the weights of the marked items. The positively weighted responses (+1) are masculine; and the negatively weighted ones (−1) are feminine, so a high positive score is highly masculine. The key was derived empirically by the authors as a result of administration of the test to several thousand male and female subjects.

Statistical results.—Data available on this test are confined to item-analysis results, based upon the responses of unclassified aviation students tested in September and October 1944 at Psychological Research Unit No. 3.

(1) *Internal consistency.*—On the basis of total masculine score, responses of the highest and lowest 27 percent of 750 unclassified aviation students were analyzed. Results of this analysis are given in table 25.5.

TABLE 25.5.—*Internal consistency data for Reaction Speed, CE451AX2, based on responses of 750 unclassified aviation students*

Type responses	N responses	Mϕ	SDϕ	Range of ϕ	
				Low	High
Masculine	188	0.10	0.14	−0.30	.54
Feminine	195	.05	.14	−.50	.39
Unscored or ambiguous	85	−.01	.11	−.40	.34

It is apparent from these data that there is little homogeneity even among responses keyed as masculine. It is also interesting to note that unscored or ambiguous responses yielded a slightly negative mean, while those scored feminine were correlated only slightly less with masculine score than were the masculine items. These facts strongly suggest the need for a new empirically constructed key for the aviation-student population.

Evaluation of Masculinity-Femininity Tests

Lack of validation data for the tests in this section leaves the results of this research inconclusive insofar as proof or disproof of the hypothesis is concerned.[4] The information approach appears to demand and warrant much more thorough research in order to secure a body of truly discriminating items. When that is accomplished, the task of proving the masculinity hypothesis still remains. One important drawback to this approach is the constant change of facts and significance of facts, which necessitates frequent revision of information items. The reaction-speed tests are vulnerable in this respect to a somewhat less degree. The fact that the masculine and the feminine items, as determined empirically by the authors, yielded only slightly different correlations with total masculinity score suggests that lack of internal consistency is a major handicap to the use of this method.

MEASURES OF CAREFULNESS

The tests in this area were conceived and designed specifically to assist in the selection of men likely to be successful in navigation. It was hypothesized that the extreme care and exactness required in navigation would be displayed in the performance of potentially successful candidates on navigational job-sample tests. The four tests in this area were constructed for the purpose of determining whether carefulness per se is related to navigation success and whether a factor of carefulness could be isolated and identified. They were administered in January 1945 for purposes of correlational analysis to 354 unclassified aviation students at Medical and Psychological Examining Unit No. 6. All the data that follow are based upon this sample.

It is important to note that the instructions to all four tests do not stress either speed or accuracy to the exclusion of the other. The instruc-

[4] Results of other research in this area are reported in ch. 21.

tions state simply, "This is a test of your ability (to plot movements on a chart; or, to plot a chart; or, to read scales) quickly and accurately."

Directional Plotting, CE455A [a]

Since the navigator's task includes a great deal of careful measuring and plotting on maps and charts, it appeared that a test measuring the speed and accuracy with which an examinee can locate points and estimate directions on a chart would be in order. The technique employed in this test differs in some respects from those employed in other plotting tests described in this section.

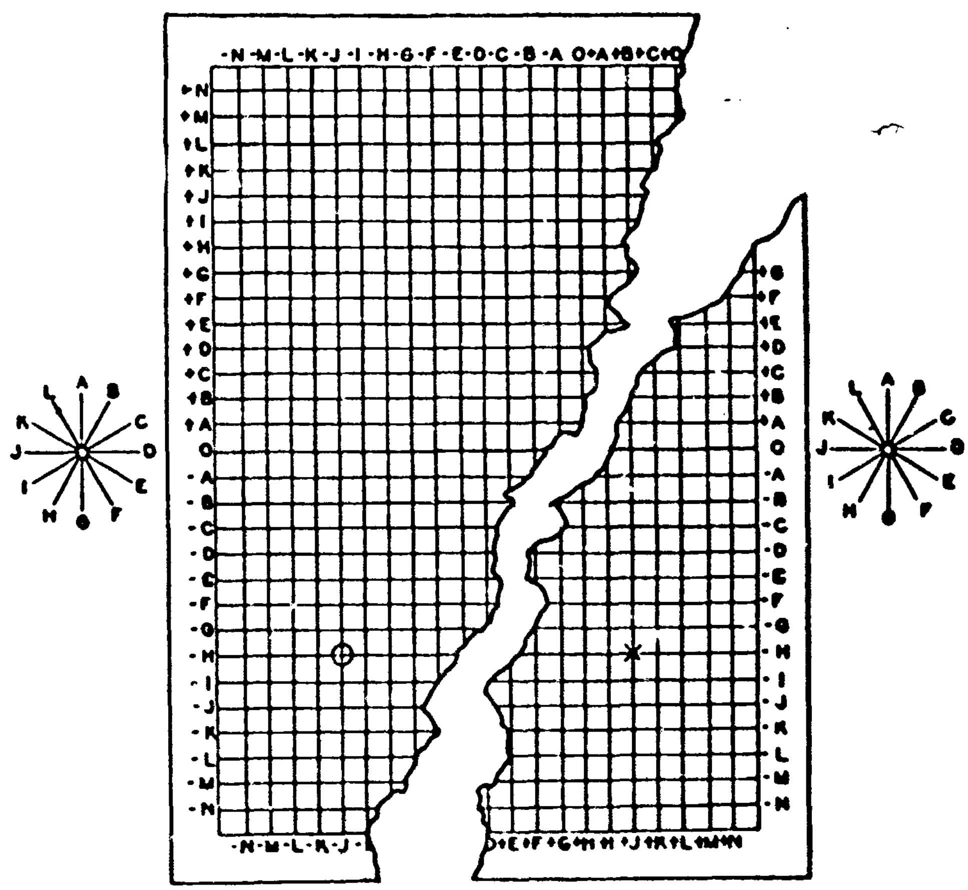


FIGURE 25.1
CHART USED IN DIRECTIONAL PLOTTING, CE455A

Description.—The chart pictured in figure 25.1 is basic to this test. The examinee is given the coordinates of two points on the chart; such as +J−H and −J−H, which are marked in figure 25.1. The examinee's task is (1) to locate the positions, but make no mark on the chart, and (2) to determine the direction of the second position from the first

[a] Developed at Psychological Research Project (navigator) and Psychological Research Unit No. 3. Chief contributors: Pvt. Loretta K. Buckley, Lt. Frank J. Dudek, Sgt. Nathan Kravetz, Lois G. Wright.

in terms of the points on the marginal diagrams. In the illustration, position −J−H is in direction J from point +J−H. If an asterisk (*) appears before the coordinates of the first position, the order is reversed, i. e., the examinee identifies the direction of the first position from the second. Thus, if the positions in figure 25.1 were written *+J−H and −J−H, the correct answer would be D rather than J.

The test booklet is printed separately and has the problems listed in the following manner:

	First point	Second point
4.	+I −I	+D −K
7.	*−K +N	−N +L

The test contains two parts, part I having 21 problems and part II having 25 problems.

(1) *Administration.*—Each examinee is supplied with a test booklet in which are inserted a chart and a standard 15-place answer sheet. Approximately 5 minutes are consumed in reading the directions and doing the sample items. Time for part I is 8 minutes, and for part II, 7 minutes.

(2) *Scoring.*—Two scores are obtained, one for total number of correct responses and one for number of errors.

Statistical results. (1) *Distribution statistics.*—Distribution data were computed for right and wrong scores separately. These data are given in table 25.6.

TABLE 25.6.—*Distribution of scores of 354 unclassified aviation students, and reliability coefficients for Directional Plotting, CE455A*

Score	M	SD	r_{tt}
Right	16.4	6.6	0.76
Wrong	9.7	4.2	.56

(2) *Reliability coefficients.*—Utilizing the raw data from which distribution constants were derived, preliminary estimates of reliability of the right and wrong scores were computed by means of Kuder-Richardson formula No. 21, and are given in table 25.6. The same sample yielded a correlation of −0.48 between right and wrong scores.

(3) *Factorial composition.*—The most significant loadings for the right scores are in the visualization (0.45), numerical (0.44), space III (0.42), spatial-relations (0.30), and psychomotor-precision (0.26) factors. The loading in the carefulness factor is only −0.03. The communality was found to be 0.76. Principal loadings for the wrongs scores (reflected) are in the visualization (0.56) and carefulness (0.41) factors. The communality was 0.50. For a fuller picture of the factorial composition of this test, see appendix B.

Complex Scale Reading, CE454A [8]

Description.—This test was designed to measure the ability to read values from scales with speed and precision.

(1) *Internal characteristics.*—The principal instrument in this test is a chart, a portion of which is shown in figure 25.2. The chart is printed on a separate sheet, while the item descriptions are printed in a booklet. The following are typical items:

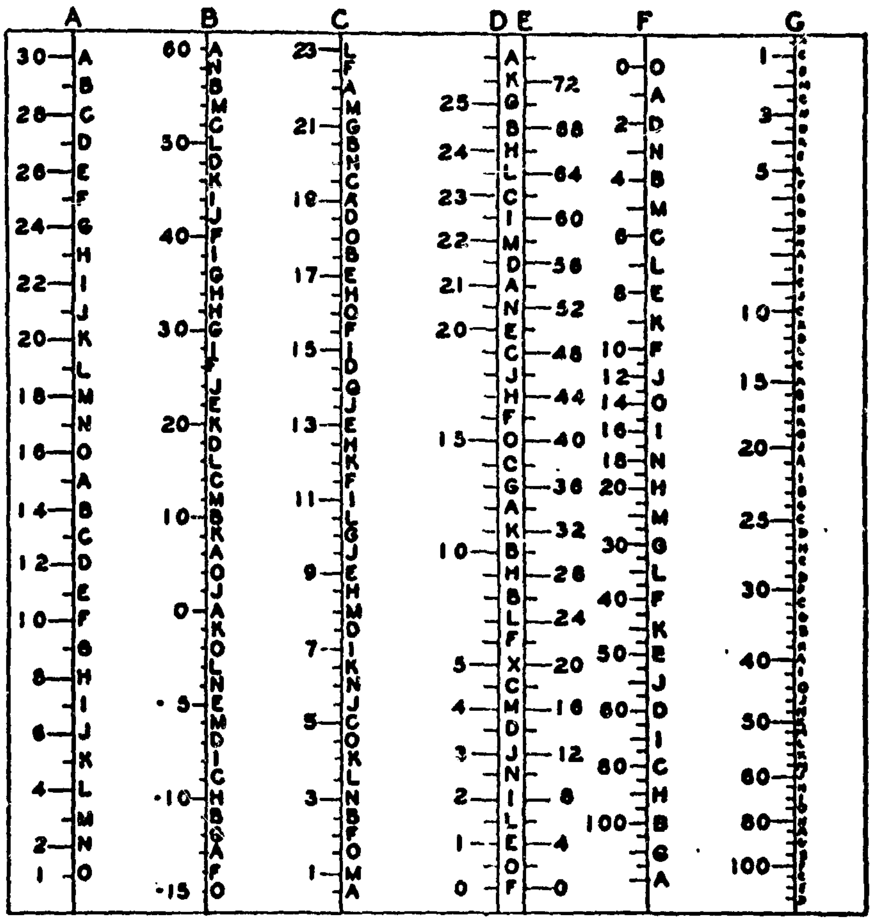


FIGURE 25.2
SCALES USED IN COMPLEX SCALE-READING TEST, CE454A

Item	*Scale values*	*Read value on scale*
5.	B40, D20	G
6.	A15, F32.5	B

The examinee's task is to locate on the chart the points listed under scale values, place a straight edge across the chart touching the two points, and read the value on the scale indicated under "Read Value on Scale." The values are read in letters, and these letters are marked on a

[8] See footnote 5.

15-place answer sheet. The test is divided into two parts. There are 21 items in part I and 25 items in part II.

(2) *Administration.*—A straight edge, a chart, an answer sheet, and a booklet are furnished to each examinee. Test directions and four sample practice problems appear on the front of the booklet. The directions require about 5 minutes. Examinees are allotted 7 minutes for working on part I and 6 minutes for part II.

(3) *Scoring.*—The two scores are total number right and total number wrong.

Statistical results. (1) *Distribution statistics.*—Distributions of right and wrong scores are given in table 25.7.

TABLE 25.7.—*Distribution of scores of 354 unclassified aviation students and reliability coefficients for Complex Scale Reading, CE4.4A*

Score	M	SD	r_{tt}
Right	15.6	4.8	0.56
Wrong	4.8	2.8	.46

(2) *Reliability coefficients.*—Preliminary estimates of reliability of scores were made by the Kuder-Richardson formula No. 21 and are given in table 25.7. Rights and wrongs correlated −0.43 on the same sample.

(3) *Factorial composition.*—The most significant loadings for the right scores are in the numerical (0.52), spatial-relations (0.33), and space III (0.32) factors. The loading in the carefulness factor is only 0.05. The communality is 0.55. The principal loading for the wrong scores (reflected) is in the carefulness (0.57) factor. The communality is 0.37. For a fuller picture of the factorial composition, see appendix B.

Plotting Test, CE452A [7]

This test and the Plotting Accuracy test, described as a variation, represent further attempts to measure functions or abilities important to the navigator.

Description.—The directions describe this test as a measure of the ability to plot movements on a chart quickly and accurately. The chart, by means of which the problems are solved, is printed on a sheet separate from the test booklet.

(1) *Internal characteristics.*—The test booklet contains the directions and a list of problems which are merely descriptions of moves on the chart. The list is divided into two parts. Part I contains problems 4 through 25, and part II, problems 26 through 50. The first three numbered problems are samples given with the directions. A reproduction of the chart, in reduced size, is presented in figure 25.3.

[7] See footnote 5.

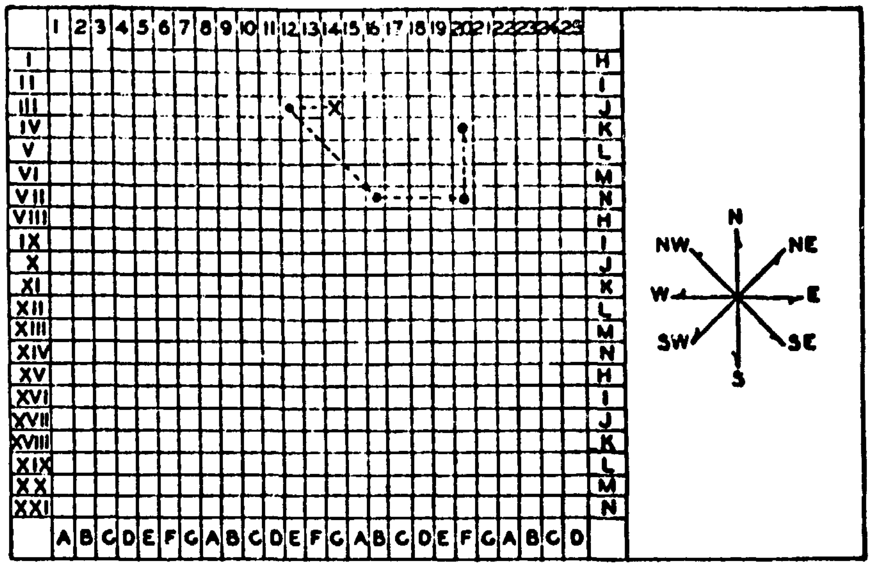


FIGURE 25.3
CHART USED IN PLOTTING TEST, CE452A

The compass rose at the right of the chart is presented in order to eliminate, insofar as possible, the influence of differential training. The problem worked out on the chart illustrates the type of item in the test and the method of solution to be employed. The square marked "X" is the starting point and is designated as III-14, the starting location being identified with reference to the left and upper marginal keys.

(2) *Administration.*—Utilizing the orientation provided by the compass rose, the examinee is directed to make (or, in the sample problem, to follow) the moves listed. The moves are to be visualized but never marked on the chart. The coordinates of the last location are the answers to the problem. In the sample, the final location is FK, the coordinates being read now from the bottom and right, in that order. Special emphasis is given to the necessity of marking both coordinates as answers. The testing time for part I is 8 minutes and for part II, 7 minutes. Reading directions and the like required approximately 6 minutes in addition.

(3) *Scoring.*—As indicated above, two letters are marked on the 15-place answer sheet for each item. Because of the method of answering, an examinee might get one letter correct even though he ended up in the wrong square. It would thus be theoretically possible for one to get at least one-half the total letters correct without doing any of the problems correctly. In order to compensate for this, the scoring formula R-W might well be applied. In the preliminary explorations, however, rights and wrongs were scored separately.

Statistical results. (1) *Distribution statistics.*—Data computed for rights and wrongs separately appear in table 25.8.

(2) *Reliability coefficients.*—The same sample yielded the preliminary reliability estimates seen in table 25.8, as found by the Kuder-Richardson formula No. 21. The rights and wrongs correlated −0.42.

(3) *Factorial composition.*—The most significant loadings for the right scores are in the numerical (0.51), space III (0.46), spatial-relations (0.25), carefulness (0.22), and psychomotor-precision (0.20) factors. The communality is 0.65. Principal loading for the wrong scores

TABLE 25.8.—*Distribution of scores of 354 unclassified aviation students and reliability coefficients for Plotting Test, CE452A*

Score	M	SD	r_{tt}
Right	45.8	11.0	0.80
Wrong	6.5	5.1	.77

(reflected) is in the carefulness (0.59) factor. The communality is 0.39. For a fuller picture of the factorial composition of this test, see appendix B.

Variation of the test. (1) *Plotting Accuracy Test, CE453A.*[8]—With few exceptions, this test is similar to the plotting test just described. The same kind of chart is employed, but three orientation compasses are shown as compared with one in the plotting test. The points on these are marked by letter rather than directions. (In each item the examinee is instructed which compass to use.) The task is similar to that in the plotting test. The factorial composition of this test is similar to that of the plotting test, see appendix B.

Carefulness Factor Analysis[9]

Administration of the carefulness tests to a sample of 354 aviation students revealed some interesting results. It was first noted that large numbers of wrong responses were made, with considerable range and variability. Although both right and wrong scores had frequently been obtained on other tests, the results of the carefulness tests first strongly suggested (1) the need for separate statistical treatment of wrong scores, and (2) the distinct possibility that error scores might be very different functionally from correct-response scores.

In a true power test, right and wrongs scores correlate −1. As the test becomes more and more speeded, the negative correlation decreases, and it becomes possible for rights and wrongs to differ factorially. Superficial evidence for these four tests indicated that the scores are quite independent.

The data.—On the basis of the evidence and theoretical concepts already outlined, it was determined to score, intercorrelate, and analyze the right and wrong scores of these tests separately. In addition to these 8 variables, 11 classification tests were also included in the matrix. The

[8] See footnote 5.
[9] Accomplished at Psychological Research Unit No. 2. Chief contributors: Capt. Lloyd G. Humphreys and Sgt. Harold H. Singer.

TABLE 25.9.—*Correlation matrix for the Carefulness Battery*[1]

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Directional Plotting (R)[a]	...	48	57	10	46	07	46	04	36	44	42	20	46	16	27	38	34	37	45
2. Directional Plotting (W)	48	...	05	28	12	20	09	14	−10	−04	19	20	14	03	10	04	36	14	17
3. Complex Scale Reading (R)	57	05	...	43	55	13	51	14	37	44	30	04	43	17	28	49	18	34	46
4. Complex Scale Reading (W)	10	28	43	...	24	39	20	34	−05	01	11	07	12	00	03	07	23	09	16
5. Plotting (R)	46	12	55	24	...	42	71	26	49	51	35	07	32	24	34	48	11	19	32
6. Plotting (W)	07	20	13	39	42	...	19	41	−12	−03	08	−01	08	05	05	05	10	08	08
7. Plotting Accuracy (W)	46	09	51	20	71	19	...	46	48	44	32	05	35	22	29	52	12	23	34
8. Plotting Accuracy (W)	04	14	14	34	26	41	46	...	−06	−04	09	−06	13	05	08	14	14	11	16
9. SAM Rotary Pursuit	36	−10	37	−05	49	−12	48	−06	...	33	38	32	18	33	12	05	21	03	−02
10. SAM Two-Hand Coordination	44	−04	44	01	51	−03	44	−04	33	...	48	30	23	27	14	−04	35	05	06
11. SAM Complex Coordination	42	19	30	11	35	08	32	09	38	48	...	32	36	35	21	05	29	16	15
12. Rudder Control	20	20	04	07	07	−01	05	−06	32	30	32	...	07	11	05	−03	27	−01	−01
13. SAM Discrimination Reaction Time	46	14	43	12	32	08	35	13	18	23	36	07	...	26	21	16	20	21	23
14. Finger Dexterity	16	03	17	00	24	05	22	05	33	27	35	11	26	...	16	14	04	04	01
15. Speed of Identification	27	10	28	03	34	05	29	08	12	14	21	05	21	16	...	11	13	11	04
16. Numerical Operations (F)	38	04	49	07	48	05	52	14	05	−04	05	−03	16	14	11	...	−16	15	35
17. Mechanical Principles	34	36	18	23	11	10	12	14	21	35	29	27	20	04	13	−16	...	31	27
18. Reading Comprehension	37	14	34	09	19	08	23	11	03	05	16	−01	21	04	11	15	31	...	50
19. Arithmetic Reasoning	45	17	46	16	32	08	34	16	−02	06	15	−01	23	01	04	35	27	50	...

[1] Decimal points omitted. Wrongs scores are reflected.
[a] for code numbers see table 25.10.

TABLE 25.10.—*Centroid factor loadings and communalities for the Carefulness Battery*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	h^2
1. Directional Plotting (R), CE455A	74	22	06	−45	−12	−23	−15	21	19	−11	22	10	17	1.07
2. Directional Plotting (W), CE455A	34	10	54	−07	15	−30	−15	05	03	−13	20	−04	−11	.63
3. Complex Scale Reading (R), CE454A	72	−19	−19	−22	−10	28	−29	−17	24	08	−08	19	10	.96
4. Complex Scale Reading (W), CE454A	36	−29	28	22	18	14	−28	−17	21	07	09	08	−14	.58
5. Plotting (R), CE452A	77	−19	−33	17	−06	−23	−13	17	−14	25	−09	−09	08	.97
6. Plotting (W), CE452A	30	−34	23	36	23	−11	−16	16	−17	16	12	22	12	.64
7. Plotting Accuracy (R), CE453A	75	−23	−32	11	−11	−21	22	08	18	−17	−14	08	−07	.93
8. Plotting Accuracy (W), CE453A	33	−39	15	33	21	−07	24	09	16	−14	−06	17	−09	.59
9. SAM Rotary Pursuit, CP410B	42	29	−39	20	−32	−11	15	−29	11	14	23	−08	21	.81
10. SAM Two-Hand Coordination, CM101A	49	33	−26	21	−30	23	−23	33	08	−15	−11	−19	14	.86
11. SAM Complex Coordination, CM701A	57	40	−07	10	10	11	04	05	−10	−12	08	−04	05	.56
12. Rudder Control, CM120B	25	41	10	17	−13	−08	−10	−19	−07	−10	09	−09	−08	.38
13. SAM Discrimination Reaction Time, CP611D	51	10	−06	−15	17	14	08	05	05	−05	−02	07	12	.38
14. Finger Dexterity, CM116A	33	23	−27	05	24	07	11	−10	−13	−10	13	04	08	.37
15. Speed of Identification, CP610A	34	06	−13	−08	13	−05	03	10	05	15	−07	−05	−06	.21
16. Numerical Operations (F), CI702B	39	−41	−27	−31	−08	−25	−07	−15	−17	−23	−11	12	−07	.70
17. Mechanical Principles, CI903B	43	29	44	11	−20	15	11	10	18	13	08	−07	−14	.64
18. Reading Comprehension, CI614H	40	−15	24	−26	−17	19	25	08	−14	10	11	05	06	.49
19. Arithmetic Reasoning, CI206C	47	−31	21	−32	−28	17	11	09	−17	−11	11	13	−10	.67

[1] Decimal points have been omitted from the centroid factor loadings.

TABLE 25.11.—*Rotated factor loadings for the Carefulness Battery*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	h²
1. Directional Plotting (R)[2]	44	−06	−03	17	26	45	11	42	−01	30	54	−06	−03	1.07
2. Directional Plotting (W)	08	−01	41	02	−03	56	01	−04	00	09	38	−06	−03	.65
3. Complex Scale Reading (R)	52	−02	05	19	17	−02	65	32	00	33	05	06	−03	.98
4. Complex Scale Reading (W)	09	−03	57	03	−07	09	47	05	08	13	01	04	[illegible]	.60
5. Plotting (R)	51	15	22	00	20	09	02	46	05	25	−07	56	01	.97
6. Plotting (W)	00	−09	59	10	01	−05	05	03	16	13	19	46	−04	.66
7. Plotting Accuracy (R)	69	15	19	02	18	03	−02	38	43	20	−06	05	01	.94
8. Plotting Accuracy (W)	22	−08	51	07	04	−02	−02	01	49	08	−07	00	−02	.57
9. SAM Rotary Pursuit	13	56	−08	00	38	−04	11	53	03	−08	06	01	02	.79
10. SAM Two-Hand Coordination	09	26	−07	09	−03	07	−05	38	−08	62	−10	−02	−05	.84
11. SAM Complex Coordination	08	35	06	12	34	24	00	14	06	46	07	00	01	.56
12. Rudder Control	−03	48	06	−03	00	27	06	09	−05	13	14	−06	−03	.36
13. SAM Discrimination Reaction Time	23	−01	03	17	35	14	14	08	12	33	06	−01	−06	.38
14. Finger Dexterity	11	26	02	00	44	−03	−01	−03	02	27	07	03	07	.36
15. Speed of Identification	20	−03	00	−03	22	20	06	13	09	16	−07	16	08	.22
16. Numerical Operations (F)	81	01	04	02	00	−10	05	−06	−01	02	16	12	−04	.72
17. Mechanical Principles	−13	18	17	34	−02	54	16	28	18	12	−03	−11	01	.65
18. Reading Comprehension	23	−06	10	60	16	20	06	06	04	−02	05	06	−04	.51
19. Arithmetic Reasoning	48	−05	13	63	−06	14	09	02	03	06	13	00	03	.70

[1] Decimal points have been omitted from the factor loadings.
[2] For code numbers see table 25.10.

list includes several apparatus classification tests, because they showed such high correlations with the carefulness tests. Descriptions of the selected classification tests appear elsewhere in this report. Intercorrelations of the variables appear in table 25.9. Intercorrelations of the carefulness tests and correlations of the carefulness tests with classification tests are based on a sample of 354 unclassified aviation students. Intercorrelations of the classification tests are based upon other comparable samples of unclassified aviation students, with a total N of 1,920. Centroid loadings appear in table 25.10 and rotated factor loadings in table 25.11. The axes of wrong scores were reversed in correlating the variables, so that low error scores, when associated with good performance in other variables, would produce positive correlations.

The factors.—Thirteen factors in all were extracted, eight of which are more or less well identifiable as genuine factors. Of the remaining five, four resulted from the fact that both right and wrong scores for any given test were derived from the same sample. The effect of this was to introduce four doublet factors, one for each carefulness test. The fifth factor is residual. Each common factor will be discussed, along with the principal tests with projections on the factor. No tests having loadings below 0.25 will be listed in the following groups.

Rotated factor I is identified by the following data:

Test No.	Test Name	Loading
16	Numerical Operations (F)	0.81
7	Plotting Accuracy (R)	.69
3	Complex Scale Reading (R)	.52
5	Plotting (R)	.51
19	Arithmetic Reasoning	.48
1	Directional Plotting (R)	.44

This is easily identified as the numerical factor, which has been isolated in other analyses. The weighted averages of the loadings of Numerical Operations (front) (0.78) and Arithmetic Reasoning (0.48) in several analyses make this identification positive. This is one of the two factors in which significant loadings appear for right scores of every one of the carefulness tests. The Plotting Accuracy and Plotting tests both involve simple counting, in addition to other functions. The Complex Scale Reading test involves points on numerical scales. This may be rationalized as numerical facility in the sense that numbers must be retained and recognized quickly. The task in the Directional Plotting test is more difficult to identify as numerical. The best explanation is probably that a sort of mental counting takes place in locating the points, even though no numbers are used in the chart.

Rotated factor II is identified by the following data:

Test No.	Test name	Loading
9	Rotary Pursuit	0.56
12	Rudder Control	.48
11	Complex Coordination	.35
10	Two-Hand Coordination	.26
14	Finger Dexterity	.26

The tests appearing on this factor clearly identify it as the psychomotor-coordination factor previously isolated. Loadings are in general agreement with those in other analyses.

Rotated factor III is identified by the following data:

Test No.	Test name	Loading
6	Plotting Test (W)	0.59
4	Complex Scale Reading (W)	.57
8	Plotting Accuracy (W)	.51
2	Directional Plotting (W)	.41

This is an entirely new factor, which appears to be uniquely characteristic of the wrong scores of the carefulness tests, at least in this matrix. Right scores of Plotting and Plotting Accuracy show slight, but probably not significant, loadings (0.22 and 0.19 respectively) with the factor. In the light of this evidence, the factor was named carefulness. Had the error scores not been analyzed, the general conclusion would have been that no new factor resembling carefulness could be found in carefulness tests. To what extent this factor is common to error scores in other tests is still to be determined.

Rotated factor IV is identified by the following data:

Test No.	Test name	Loading
19	Arithmetic Reasoning	0.63
18	Reading Comprehension	.60
17	Mechanical Principles	.34

This variable is difficult to reconcile with findings of other analyses as a single factor. The loading of Arithmetic Reasoning suggests the general-reasoning factor, but the loading for this factor is larger than the weighted average (0.47) for several other analyses. Reading Comprehension has a much smaller average loading on the factor (0.19) than it has in this analysis, while the weighted average loading for Mechanical Principles (0.34) is identical with that obtained here. Weighted averages of loadings in the verbal factor for the same three tests are 0.27, 0.60, and 0.03, respectively. Combining mean weighted-average variances in the general-reasoning factor with those in the verbal factor gives totals of 0.54, 0.63, and 0.34 respectively for the tests appearing significantly on this variable in this analysis. The evidence cited seems to indicate that the factor is a rough combination of the general-reasoning and verbal factors which are identified separately in other analyses.

Rotated factor V is identified by the following data:

Test No.	Test name	Loading
14	Finger Dexterity	0.44
9	Rotary Pursuit	.38
13	Discrimination Reaction Time	.35
11	Complex Coordination	.34
1	Directional Plotting (R)	.26

This is apparently the same factor that was previously identified (in the analysis of the December 1942 battery; see ch. 28) as the psychomotor-precision factor. It is not difficult to see the content, as described by the term psychomotor precision, in most of these tests.

Rotated factor VI is identified by the following data:

Test No.	Test name	Loading
2	Directional Plotting (W)	0.56
17	Mechanical Principles	.54
1	Directional Plotting (R)	.45
12	Rudder Control	.27

The loading of Mechanical Principles on this factor, together with the loadings of other tests, indicates that this is the factor identified in earlier analyses as visualization. The fact that Directional Plotting, rights and wrongs, are both relatively high in the factor is revealing. The task in this test involves determining direction on a chart in terms of the points on a compass rose that is located on the margin of the paper. The ability to visualize the compass rose on the plot itself would obviously assist in getting a high right score. Achievement of a low negative score is apparently even more a function of this ability. A formula score for this test would yield a much higher loading for this factor than has been found in any test yet analyzed.

It is probably significant that the loading in the visualization factor was found in previous analyses to be inversely proportional to the speed demanded in the task. Interpreting the loading of wrong scores on this factor in the light of this evidence suggests that the low wrongs scores are made by the more deliberate but meticulous individuals who employ a rather exact type of visualization.

The moderately low loading of Rudder Control (0.27) on this factor may be due to the necessity of visualizing the direction and amount of movement to be made in order to correct the imbalance. The kinesthetic sense is undoubtedly important in this test, but it must apparently be supplemented by visual imagery of the relative position of apparatus and target. Failing the latter, the examinee may be engaging in continual trial and error.

Rotated factors VII, IX, XI, and XIII were anticipated and rotations were made first to those positions in order to facilitate the rotation of the other reference axes. It will be noted that each of these is identified only by the right and wrong scores of one of the carefulness tests. These factors are best explained as resulting from (1) true nonerror specific variance and (2) the correlation of errors in right and wrong scores obtained from the same test and sample.

Rotated factor VIII is identified by the following data:

Test No.	Test name	Loading
10	Two-Hand Coordination	0.58
9	Rotary Pursuit	.53
5	Plotting (R)	.46
1	Directional Plotting (R)	.42
7	Plotting Accuracy (R)	.38
3	Complex Scale Reading (R)	.32
17	Mechanical Principles	.28

The identity of this factor is difficult to establish, but the combination of tests appearing with the factor indicates that some sort of spatial ability is the common element. Because the Two-Hand Coordination test was highest, the factor was named space III (two-hand coordination) until a more definite description can be achieved. A characteristic of the factor may be tentatively hypothesized as "spatial reference." Change in location and speed are important in Two-Hand Coordination; change of location and possibly distance are factors in Rotary Pursuit; distance and direction are factors in the Plotting, Directional Plotting, and Plotting Accuracy tests; Complex Scale Reading involves direction and location; and some problems in the Mechanical Principles test involve direction and possible change in distance. Although the loading for the latter test may be due to the absence of the mechanical-experience factor in this analysis, all these tests appear to involve a spatial reference factor. It cannot at present be identified with space II (hands), in spite of the fact that the leading test in the list here is Two-Hand Coordination, in which a right-left space discrimination is very apparent. This test appeared in the same analysis with the Hands test—the Integration Battery analysis (see ch. 10)—without showing any space communality over and above that of space I. More information should be secured before positive identification of factor space III is made.

Rotated factor X is identified by the following data:

Test No.	Test name	Loading
10	Two-Hand Coordination	0.62
11	Complex Coordination	.46
13	Discrimination Reaction Time	.33
3	Complex Scale Reading (R)	.33
1	Directional Plotting (R)	.30
14	Finger Dexterity	.27
5	Plotting (R)	.25

The factors bear considerable resemblance to the spatial-relations or space I factor identified in other analyses. Weighted average loadings

703120—47—45

from several analyses on the spatial-relations factor for the Complex Coordination, Two-Hand Coordination, Discrimination Reaction Time, and Finger Dexterity tests are 0.49, 0.41, 0.42, 0.17 respectively. The much higher loading of Two-Hand Coordination in this analysis is difficult to rationalize. The loadings of carefulness tests (Plotting Accuracy has a loading of 0.20) indicate that this factor identifies another characteristic in which the four tests are similar. In terms of the description of the factor, the common element appears to be the locating and relating of fixed and spatially separated points. With the description of this factor involvement in the carefulness tests (right scores), we have evidence that all of them resemble the tests of Dial Reading and Table Reading functionally, with loadings in numerical and space I quite comparable.

Rotated factor XIII appears to be a true residual.

Conclusions.—This analysis produced some new and useful information. It is interesting to note that, for the first time, an entire group of printed tests proved to have more in common with apparatus tests than with other printed tests. This fact gives considerable encouragement to proponents of the belief that many factors appearing in apparatus tasks can be duplicated in printed tests. It is also noteworthy that this analysis produced more complete identification of the apparatus tests than any one analysis had previously produced.

Probably the most significant result of this analysis is the discovery that analysis of wrong scores brought to light an entirely new factor. This fact may have important implications for future factor analysis and test construction. It may be assumed safely that, if correlations of right and wrong scores are not too high, a fuller picture of the true functions measured by a test can be obtained by analyzing the scores separately than by analyzing formula scores. The results also imply that many an error has possibly been committed by combining right and wrong scores in the same formula. Unless the two are factorially similar the result may be very different than had been intended by the test maker. The finding actually opens up a whole area of research on the use and weighting of error scores in printed tests. In this connection, it is in order to suggest that if right and wrong scores from a test are considered in factor analysis or in composite predictions, they should be derived from separate forms in order to avoid spurious correlations.

Although more investigation should be made to confirm the findings of this analysis, the discovery of a carefulness factor is certainly significant. Nothing is known as yet concerning the validity of the factor for air-crew selection, but the uniqueness of the wrong scores which identify the factor suggest that it would add much to the classification stanines even though validity of the factor is not high. The new space factor (space III) is also a discovery which extends to some degree the knowledge of apparatus tests. If the factor proves valid, it will assist further in accounting for the total validity of these tests.

Evaluation of Carefulness Tests

Most of the results of the research in this area are covered in the concluding statements regarding the factorial study. Nothing is known, as yet, concerning the validity of these tests for any air-crew positions. It appears logical to assume, however, that wrong scores, at least, will prove useful in selection for clerical-type work and possibly for navigation.

As a characteristic of temperament, little more can be said concerning carefulness than was said in the discussion of the rationale that stimulated preparation of the tests. It may be added, however, that we know that these tests identify a factor of some kind, and that the factor is presumed to be carefulness which, as presently conceived, is a trait of temperament. How far the trait is generalized to other test and job situations is yet to be determined.

MEASURES OF FEAR AND TENSION

The purpose of the tests in this section was to identify those men who are less likely to succeed in the flying situation by reason of fear of physical danger. Traits of temperament, such as fearfulness, occur in varying degrees of intensity and with varying degrees of generality in different individuals. Such traits manifest themselves in the overt behavior, mental attitudes, and other implicit emotional reactions. Many expressions of fear are difficult to evaluate objectively. Among those that may be subject to objective measurement are verbally expressed opinions.

Survey of Aviator Opinion, CE604B [10]

It appeared logical to assume that attitudes of individuals toward danger would be reflected in their opinions regarding aviation practices, construction of planes, methods of training, and the like. These expressions would be valid indices of attitudes only if no irrelevant motivation for certain responses exists. This test is an outgrowth of part II of the original Biographical Data Blank, CE602A (see ch. 27).

One aspect of motivation contributing to invalidity of test responses is the general social unacceptability of the exhibition of fear. Because of this, it seemed advisable to present material in this survey in such a way that the examinee would regard none of the available responses as meriting social disapproval. Careful analysis of the responses should make it possible to distribute or rank individuals according to the proportion of responses identified as associated with or symptomatic of fear of physical danger.

In line with these requirements, it was determined to construct a survey which would explore the cadet's opinions about methods, policies,

[10] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. S. W. Cook, Col. J. P. Guilford, Capt. L. G. Humphreys, and Lt. David H. Jenkins.

equipment, and the like, employed in military aviation. An attempt was made to describe the situations in such a manner that the examinee would respond in the role of an emotionally uninterested observer. Presumably this approach would minimize equivocation, since opinions so expressed should appear to the examinee to be devoid of social implications.

Description (1) *Internal Characteristics.*—This form consists of 52 statements about flying. The student is required to indicate his attitude toward each statement on the following 5-point scale:

A. Strongly agree.
B. Agree.
C. Undecided or have no opinion.
D. Disagree.
E. Strongly disagree.

He enters his reaction to each statement opposite the appropriate space on the separate answer sheet. The following statements are typical of those appearing in the test:

In building military planes it is well to sacrifice speed in order to add more armor.

A course of parachute jumping should be given every student flyer before he begins flying training.

Aerobatics should be reduced to a minimum in training.

The quality of the plane is the biggest factor determining a pilot's success.

(2) *Administration.*—The Survey is practically self-administering. Brief directions on responding and marking the answer sheet suffice. Although it was desired that all examinees respond to every statement, a time limit of 12½ minutes was established. Students are urged to record their first reactions, rather than those resulting from long and careful consideration. Most examinees finish in the allotted time.

(3) *Scoring.*—Owing to the nature of the material involved, responses do not fall into right and wrong categories. Either of two methods of scoring can be utilized in such a case. Either an a priori key can be made, or a key developed from validation of the responses to the items. The latter method was employed. A relatively large sample to which the Survey was administered was divided into two equal parts (odds vs. evens). Response validities against the criterion of graduation or elimination from primary pilot training for the two groups were calculated separately. On the basis of these validities, keys were constructed. The key of the odd sample was used in scoring the even sample, and the even-sample key was used in scoring the odd sample. This was done in order to avoid the "bootstrap" effect of scoring a sample by means of an empirical key derived from the same sample. As a result of the information obtained from this response validation, each response was scored either plus 1, minus 1, or zero. The scoring formula is the algebraic sum of the response weights. A constant of 20 was added to all scores in or-

der to eliminate negative scores. The key derived from sample I contains 59 responses scored plus 1 and 52 responses scored minus 1. The key derived from sample II contains 52 responses scored plus 1 and 59 responses scored minus 1.

Statistical results. (1) *Distribution statistics.*—The two samples on which the empirical scoring keys were based yielded the distributon constants given in table 25.12.

TABLE 25.12.—*Distribution of scores on Survey of Aviator Opinion, CE604B, for samples of pilot trainees*[1]

Sample	N	M	SD
I	683	32.3	6.1
II	685	25.4	5.3

[1] In class 44F. Tested at Psychological Research Unit No. 3.

(2) *Test validity.*—The two samples previously referred to yielded the data in table 25.13 based on a dichotomy of primary, basic, and advanced eliminees plus those rated as "below-average" [11] in primary training versus all others in the samples who were graduated from advanced. This criterion was adopted in preference to the usual primary pass-fail criterion because the elimination rate was low at this time and only by some such device could the low group be made sufficiently large to make validity results stable.

TABLE 25.13.—*Validity of Survey of Aviator Opinion, CE604B, for prediction of success in pilot training*

N_t	P_s	M_s	M_t	SD_t	r_{bis}	$_c r_{bis}$[1]
[2]683	0.74	32.46	31.82	6.09	0.06	0.07
[2]685	.74	25.58	24.84	5.30	.08	.09

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 44F. Tested at Psychological Research Unit No. 3.

(3) *Item validity.*—Validities of all responses were computed for the two samples separated according to the dichotomy already described. Another sample was validated against the pass-fail criterion in primary training. Means and standard deviations of phi values are given in table 25.14. The most valid response of the five, without regard to sign, is used in each item as the basis on which these statistics are given.

Further knowledge concerning the usefulness of this instrument was sought in a study of pilot trainee performance. Three items of the Rating Cadet Performance SA-T2 [12] scale were used as criteria against which the items of the Survey of Aviator Opinion were validated. The three items are: Item No. 5, relaxation in flight: ability to relax during flight (freedom from tenseness); Item No. 9, social confidence: ease

[11] These ratings were secured from pilot proficiency cards.
[12] A rating scale used in a landing-study project. This scale is not described in this report.

TABLE 25.14.—*Validity phis based on most valid response to items of Survey of Aviator Opinion, CE604B*

Group	Criterion	N_t	r_e	$M\phi$	$SD\phi$	Range of ϕ	
						Low	High
Pilots in primary training	Graduation-elimination	[1]675	0.89	0.09	0.04	0.03	0.20
Pilots through advanced training (Group I)	Below-average primary trainees, primary, basic, and advanced eliminees vs. all other graduates.	[1]677	.74	.06	.03	.01	.12
Pilots through advanced training (Group II)	Below-average primary trainees, primary, basic, and advanced eliminees vs. all other graduates.	[1]677	.74	.07	.03	.02	.15

[1] In class 44F. Tested at Psychological Research Unit No. 3.

with which he approaches you to ask questions and express his opinion; Item No. 10, potential ability: promise he shows for future success as a pilot.

For purposes of the analyses, the results of which are given in table 25.15, the group of 140 was split into the high and low 50 percent on each of the three criterion ratings described. The ratings were made on an eight-point scale with the extremes described as Least . . . and Most . . . . and the center as middle of the group. Responses to the items in the Survey of Aviator Opinion varied greatly in validity against these criteria, and the relationships between validities for the various criteria were not high. The means and standard deviations of the most valid responses to items were very similar, however, as indicated by the data in table 25.15. This similarity is probably due in large part to the intercorrelations among the criteria.

TABLE 25.15.—*Validity phis based on most valid response to items of Survey of Aviator Opinion, CE604B, using rating criteria for pilots in primary training*

Criterion	N_t	$M\phi$	$SD\phi$	Range of ϕ	
				Low	High
Item No. 5 in Rating Cadet Performance	140	0.10	0.05	0.04	0.21
Item No. 9 in Rating Cadet Performance	140	.10	.05	.02	.22
Item No. 10 in Rating Cadet Performance	140	.11	.04	.03	.23

The positive phi values obtained indicate correlation of Survey of Aviator Opinion item responses with desirable ratings in the three traits. It is significant that the ranges of phi values were not great. Owing partly to this fact, the rank orders of response validities based on the various criteria did not agree closely.

Evaluation.—As an instrument for measuring attitudes of fear and caution, this survey may be useful. The validities reported do not seriously cast doubt upon the potential usefulness of the instrument, since the criteria employed were inappropriate. Criteria involving actual manifestations of fear might yield quite different results. It is likely, however, that reliable diagnosis and prognosis can be made only for those who vary extremely from the norm of such an instrument. Accurate prediction, even for these, appears to depend upon the extent to which the aspect of the social unacceptability of the fear response can be masked. To whatever extent it is possible to secure accurate responses from the chronically fearful and over-cautious, it is probable that the technique employed in Survey of Aviator Opinion is useful.

Variations of the test

(1) Survey of Aviator Opinion, CE601A[16]—This is the first form of the Survey, which contains 45 items similar to those described under

[16] See footnote 10.

CE604B. The original nucleus of 20 items appeared in an early form of Biographical Data. Form CE604A was administered to 90 superior fighter pilots and 47 superior heavy-bomber pilots in class 44B, and an item analysis was made. Out of a total of 225 possible responses in the test, 193 were selected by 5 percent or more of the group. Of this number, 38 responses yielded phi values of 0.15 or greater with the fighter-bomber dichotomy. Of the 45 items, 20 yielded phi values of this magnitude for one or more responses.

(2) Survey of Aviator Opinion, CE604C[14]—On the basis of validation against the Rating Cadet Preference items, a careful inspection was made to determine what types of items were most discriminating. Certain characteristics of opinion were common to those who were rated as lacking in confidence. This group tended to favor more thorough instruction in ground school and special phases of flying. They did not favor training that is dangerous. They favored safety precautions in flying, safer planes, and protection of cadets from off-duty danger by rules against motorcycling and the like. In general this group felt that fear and tenseness are unimportant and can be overcome. Consistent with this is the belief also expressed that relaxed pilots are not necessarily good and that slow learners should be given special help.

In line with these findings, a new form of Survey of Aviator Opinion was constructed, containing 60 items. More items of the type that showed discrimination in the B form were constructed, and items of nondiscriminating character were omitted. This form was administered for validation but data were not available at the time this was written.

Stress Resolution, CE441A[15]

It has been hypothesized that normal individuals succumb to combat fatigue because of the abnormal stresses imposed upon them by combat conditions. The use of the term "normal," in this connection, is inexact, but the fact remains that individuals display a wide range of reactions to stress situations, and no valid method has been found for predicting these reactions.

The devisers of this test set forth the hypothesis that individuals may be placed in three classes according to their reactions to stress situations. In the first group are those who look upon the opportunities that a stress situation has to offer as being more important than its threat of failure or loss. This group contains the rough-and-ready individual who is always willing to take a chance.

In the second group are those individuals who see the threat of loss or failure as more important than the chance of success. These individuals are conservative and avoid, whenever possible, the necessity of taking a chance.

[14] Developed at Psychological Research Unit No. 3. Chief contributor: Cpl. Harold H. Kelley.
[15] Developed at Psychological Research Unit No. 1. Chief contributor: T/Sgt. Ralph H. Turner.

The third group consists of those middle-of-the-road persons who have no strong predilection to either seek or shun the chance situation. The direction of their motivation fluctuates about a mean, where possible advantage in the chance situation is evenly balanced in their thinking against possible disadvantage.

The hypothesis held that the second group should be more susceptible to combat fatigue by reason of the conflict created by the military situation. The military mores encourage and laud the taking of dangerous and often costly risks, while the bent of the individual is toward careful, conservative conduct. Ambivalence results, since the individual wants to do what is socially approved but is emotionally unsuited to such action.

Description.—The designers of this test attempted to describe situations in such a manner that the examinee would have occasion to demonstrate the extent to which his choices are influenced by considerations of absolute security in preference to precarious opportunity.

(1) *Internal characteristics.*—Part I of the test consists of five problems, each of which includes seven items. The following sample illustrates the method of presentation and content of this part.

You have just been assigned to a new job as an assistant instructor in assembling and disassembling machine guns. You have seen the others work with the guns but have not had a chance to actually handle this type of gun yourself. You receive a phone call asking for someone to demonstrate the gun. You do not know when the regular instructor will return.

In each of the following circumstances if you would go ahead and try to demonstrate the gun yourself, even though you know you are not prepared, blacken the space under A. If you are not sure what you would do, blacken B. If you would try to get out of it by asking them to wait until the regular instructor returned, blacken C. The demonstration is to be before:

22. A group of buck privates.
23. A group of commissioned officers.
24. A group of noncommissioned officers.

Part II contains ten statements of opinions or principles regarding luck and chance. An excerpt from this part, including directions for answering on a five-point scale, follows:

If you strongly agree with one of the following statements, blacken the space beneath A. If you agree, blacken space B. If you aren't sure, blacken C. If you disagree, blacken D. If you strongly disagree, blacken E.

In everyday life situations:

36. A person who trusts to luck will be more successful than one who doesn't.
37. Taking a chance is a bad thing.
38. A person should leave well enough alone.
39. If a person trusts to luck, he is not using his head.
40. When a person takes a chance, he has everything to lose and nothing to gain.

Part III consists of 12 information items. Ten of these items contain fictitious names or information, so no right answer is possible. For face validity, two are authentic; e. g., item 46 in the following test excerpt:

In the following examination, if you believe the correct answer is A, blacken the space under A. If you believe the correct answer is B, blacken the space under B. If you are not sure of the correct answer, and do not wish to guess, blacken the space under C.

You are given 10 points to start with. For each correct answer you will receive an additional point. For each incorrect answer you will lose a point. If you are not sure and do not wish to guess, your score will not be affected because marking space C does not influence your score.

46. The experiments of Wilbur and Orville Wright were carried out at:
A. Kitty Hawk. B. Pelican Bay. C. Not sure.

47. The term "dihedral" was first used in a book published by:
A. Captain H. A. Smith. B. Captain W. J. Bowles. C. Not sure.

(2) *Administration.*—Initial test instructions are simple and short. As indicated in the sample items given, the method of marking is explained at the beginning of each problem or section.

(3) *Scoring.*—In the absence of an empirical key or weighting system, subjectively determined weights were assigned to all responses on a five-point scale. Responses that indicated greatest desire for security were given weights of 1, while those indicating least consideration for security were assigned weights of 5. Intermediate responses received weights of 2, 3, or 4.

Results and evaluation.—This test was administered to a sample of 1,087 unclassified aviation students in July and August 1944 at Psychological Research Unit No. 1. Separate scores for the three parts were derived and intercorrelated. Part I and part II showed considerable correlation, but part I–part III and part II–part III correlations were not significantly greater than zero. Unfortunately, data are not available on the validity of the test for prediction of air-crew success. An even more interesting study, the validation of prediction of susceptibility to combat fatigue, should be done. Only if these or some similar data are available, can the degree of correctness of the original hypothesis be determined.

Evaluation of Measures of Fear and Tension

Usefulness of expressions of opinion and attitude, as employed in these tests, appears to be limited. Results obtained from administration of the Stress Resolution test and other similar tests led to an appreciation of the limitations of subjectively derived scoring keys. Low correlations with training criteria achieved by the Survey of Aviator Opinion may indicate that responses do not have the same significance for all examinees. The validity and reliability of interpretation of these responses by both examinees and psychologists seem, therefore, to be especially significant in this type of test.

If valid interpretation of responses were achieved, however, the question of the significance of the response for air-crew success still remains unanswered. If specific fears are predictive of failure in training, a measure such as Survey of Aviator Opinion might be effective in identifying those likely to fail. If, on the other hand, the generality only of the fear response is predictive, more adequate instruments would no doubt be required. It appears, then, that fear and tension, either general or specific, may be related to air-crew success, but that analysis of results obtained from the tests in this area reveals no conclusive evidence.

MEASURES OF CONFIDENCE

Field studies, as well as casual observation by both psychologists and laymen, indicate that important differences exist in the attitudes of individuals as they attack new problems or tasks. Certain individuals enter upon such new experiences with zest and confidence, while others display considerable trepidation. It is assumed that between these extremes lies the majority of individuals who display less marked reaction to coping with new situations. Although empirical evidence is not available, it appears from observation that individuals tend to establish a uniform pattern of reaction toward such new problem situations. These facts suggest that a measure of the confidence with which individuals approach tasks might be useful in the selection of trainees.

Indices of Self Confidence, CE427A [16]

It was hypothesized that confidence should be measured, not in terms of the excellence of performance forecast by the examinee, but rather in terms of the extent to which the forecast differs from actual performance.

It was hypothesized that such a measurement would bear some relationship to air-crew training. In general, it was supposed that the more realistic individuals would prove more capable in air-crew positions. It was felt that those erring greatly in their prediction of performance would tend to make similar errors of judgment in flying situations, which would result, perhaps, in low efficiency, limited success, or even in extreme instances, death to themselves and others.

It was decided that prediction of psychomotor scores could be used. Since the candidates generally are unfamiliar with the apparatus, the factors of experience, learning, etc., would largely be avoided. Of further advantage would be the fact that little additional time and no additional tasks are required in obtaining scores by this method.

Description.—As suggested in the previous paragraphs, this measure is not a test in the usual sense but is rather an indication of the attitude toward and evaluation of new tasks.

(1) *Internal characteristics.*—A 10-point scale for rating performance was constructed, ranging from 9–10, very good, to 1–2, very poor.

[16] Developed at Psychological Research Unit No. 1. Chief contributor: Lt. George S. Klein.

The midpoint is average, point 5 is just below, point 6, just above average, poor is 3–4, and good is 7–8.

(2) *Administration.*—Two rating scales of the type described were provided each student for each of the six psychomotor tests in the classification battery. Upon entering the testing room, the candidate was required to indicate on one of the scales the level of expected performance in the test he was about to take. After taking the test and without knowledge of the score obtained, the candidate was required to indicate on the second scale what he thought his level of performance had been. While this second rating was being made, the candidate did not have opportunity to refer to his first rating for purposes of comparison. He took the test simultaneously with three other students in the same room and may have gained some impression of his relative success. His performance and relative status on earlier tests—insofar as he could appreciate it—also may have had some bearing on his predictions in later tests.

(3) *Scoring.*—Difference scores were obtained, based upon the absolute magnitude of the discrepancies that existed between each estimate —estimate 1 (pre-test) and 2 (post-test)—and the actual performance. In order to make comparisons of ratings and performance possible, standard-score norms for the points on the rating scale were established. Test-performance standard-score norms were also computed. Difference scores for pre-test estimates and for post-test estimates were transmuted into a nine-point distribution, a scale value of one indicating little or no deviation from the estimate and nine indicating extreme deviation. These 12 scores (2 for each of 6 psychomotor tests) were used for validation purposes.

Statistical results.—Statistical results for this test are confined almost entirely to small-sample validation of the difference scores just described. The samples were tested in February 1944 at Psychological Research Unit No. 1.

(1) *Test validity.*—Both pre-test and post-test difference scores were validated against the pass-fail criterion in primary pilot training. Of interest also are the correlations obtained between the difference scores and standard scores on the psychomotor tests. These data are given in table 25.16.

Evaluation.—Several tentative conclusions can be drawn from the data presented in table 25.16. For this sample, it is apparent that eliminees are on the average less accurate in both predicting and evaluating their performances in the psychomotor tests used. This fact may suggest that those successful in training are more realistic with respect to their abilities. With the exception of aiming stress, the partial correlation (holding test-score constant) between difference score and criterion was larger for the post-test than for the pre-test ratings. This fact suggests that the characteristic of confidence takes on more significance as the individual becomes better oriented in the field of performance. The

TABLE 25.16.—*Correlations of difference scores on Indices of Self-Confidence, CE427A, with various criteria, based upon a sample of pilots*

Psychomotor Test	Score[1]	N_d	r_d	M_p[2]	M_r	SD_d	r_{12}[3]	r_{13}	r_{23}[4]	$r_{13.2}$[5]
Rotary Pursuit, CM803A	I	268	0.90	4.80	5.89	1.91	—0.57	—0.30	0.25	0.20
	II	271	.89	4.82	5.83	1.90	—.54	—.28		.21
Two-Hand Coordination, CM101A	I	271	.89	4.71	5.76	1.71	—.66	—.32	.35	.13
	II	271	.89	4.61	5.97	1.90	—.58	—.37		.22
Discrimination Reaction, Time, CP611D	I	268	.90	4.85	5.39	1.83	—.48	—.15	.26	.03
	II	267	.90	4.64	5.68	1.80	—.35	—.30		.23
Aiming Stress, CE211A	I	271	.89	4.94	5.66	1.95	—.72	—.19	.11	.16
	II	269	.90	4.78	5.29	1.85	—.55	—.14		.10
Finger Dexterity, CM116A	I	267	.89	4.79	5.35	1.89	—.52	—.15	.13	.10
	II	271	.89	4.72	5.52	1.83	—.34	—.23		.21
Complex Coordination, CM701A	I	268	.89	4.88	6.14	1.99	—.59	—.33	.39	.13
	II	271	.89	4.81	5.97	1.81	—.56	—.33		.14

[1] I pertains to pre-test rating; II pertains to post-test rating.
[2] M_p, M_r, and SD_d refer to scaled difference scores between rating and performance.
[3] 1, Self-confidence difference scores; 2, standard scores on indicated psychomotor test; 3, pass-fail criterion in primary pilot school.
[4] These data not based on this sample, but are typical validity coefficients for the psychomotor tests.
[5] A correlation of 0.20 is significant at the 5 percent level; 0.27 significant at the 1 percent level.

realistic individual acquires a better basis for his judgment, while the unrealistic person probably changes little as a result of his added knowledge. The magnitude of the validity figures with performance scores partialed out suggests that further investigation should be made, although none of the validities are significant at the 1 percent level and only five are significant at the 5 percent level. On the basis of the preliminary data presented, difference scores obtained in this manner would have added considerably to the predictive value of the composite pilot-aptitude score.

Self-Crediting Mental Abilities, CE429A [17]

As the title indicates, this test was designed to measure confidence in mental rather than physical abilities.

Description.—This test consists of five parts, in which various types of tasks are provided. The content of the items is utilized only as a means of obtaining confidence scores.

(1) *Internal characteristics.*—Part I of the test presents 12 information items with four alternative answers to each and three alternative responses to indicate the confidence with which the examinee answers each question. The following item is typical of those in this part:

A barometer measures:
A. Air pressure.
B. Distance.
C. Electricity.
D. Time.

M. Certainly correct.
N. Probably correct.
O. Doubtful.

Part II contains 12 five-alternative items calling for logical selection. The following is a sample of items in this part:

An official always has:
A. A badge.
B. Duties.
C. Rights.
D. A salary.
E. A uniform.

M. Certainly correct.
N. Probably correct.
O. Doubtful.

In answering the items in part II, the examinee is required to mark the two correct answers.

Part III contains 12 items in which the examinee must indicate which one of 5 alternatives does not belong in the list because it is unlike the others. The following sample is typical:

[17] Developed at Psychological Research Unit No. 1. Chief contributors: Pfc. Vernon W. Grant, Lt. Llewellyn N. Wiley.

A. Democrat.
B. Methodist.
C. Republican.
D. Tory.
E. Whig.

M. Certainly correct.
N. Probably correct.
O. Doubtful.

Part IV contains 12 four-alternative analogy items. The following item is typical:

Seldom is to never as little is to:

A. Small.
B. None.
C. Large.
D. Often.

M. Certainly correct.
N. Probably correct.
O. Doubtful.

Part V contains 12 four-alternative number series items similar to the following:

16 17 15 18 14 19 ....

A. 13 21.
B. 13 23.
C. 13 20.
D. 12 20.

M. Certainly correct.
N. Probably correct.
O. Doubtful.

The examinee's task is to select the numbers that will carry on the series in the sequence established by the numbers listed in the problem.

(2) *Administration.*—The examinee is instructed to answer an item and then indicate the strength of his confidence in the correctness of the answer by filling in space M, N, or O. Fifteen-place answer sheets are employed for the test. The examinee is informed that if he answers M—Certainly correct, he will receive 3 points credit if the answer is correct, but will be penalized 3 points if it is wrong. N—Probably correct is weighted 2, and O—Doubtful, 1.

(3) *Scoring.*—The scores actually used in evaluating results of the test were based upon the number of M, N, and O ratings. In this way it was intended that absolute knowledge, ability, and the like would be eliminated from the score.

Results and evaluation.—Preliminary analysis of the scores revealed that approximately 85 percent of the responses were in the M, or most confident, category. The O, or doubtful, category was marked in only approximately 5 percent of the responses. These data were interpreted as indicating that the material was much too easy. The 15 percent selecting less than the most confident rating was obviously too small a proportion, and the effective range of confidence too limited to allow for reliable validation of the hypothesis. If more work is done on this test, the categories of assurance should probably be revised, adding a still more positive statement of confidence and rewording other statements. The list might then read as follows:

L. Unquestionably correct.
M. Almost certainly correct.
N. Probably correct.
O. Possibly correct.

Although the intervals in this scale are not equal, each indicates a different degree of confidence. It is possible that, by the use of a scale such as that suggested, a better distribution of expressions of confidence would be achieved.

Quantitative Estimation, CE440A [18]

It was reported that failure in flying training is frequently caused by anxiety, as expressed in lack of confidence, indecisiveness, and in other symptoms in this area. The habitual behavior of the individual in making decisions should reveal some information about his tendency toward anxiety. It was hypothesized that if the individual were given some simple task, of the five-alternative multiple-choice variety, in which correct answers could be estimated only, the tendency to anxiety and insecurity would manifest itself in indecision and lack of confidence in his answers. This might be measured by having the examinee rate his confidence in his answers in some way.

Description.—It was decided that the desired measure could best be obtained by providing for several choices by the examinee. The number of choices or guesses he took should, therefore, be a measure of his sureness of the right answer.

(1) *Internal characteristics.*—This test consists of three parts. Each item in the test has three numbers, so the examinee will have three spaces on the answer sheet and can make three guesses if he desires. Part I contains 30 items in which the examinee is required to select the correct proportions of familiar objects. The samples in figure 25.4 are typical of the items in part I.

Part II consists of 30 items in which the examinee is required to select the figure which has the largest or the smallest area from a group of five figures of various shapes. Figure 25.5 gives items typical of part II.

Part III contains 15 items in which the examinee is required to select the answer that correctly describes the size, weight, capacity, or the like, of familiar objects. Following are samples of the items in this part:

190-191-192.—The number of regulation baseballs which would weigh 5 lbs. is:
(1) 16 (2) 12 (3) 8 (4) 10 (5) 20.

[18] Developed at Psychological Research Unit No. 1. Chief contributor: T/Sgt. Louis Delman.

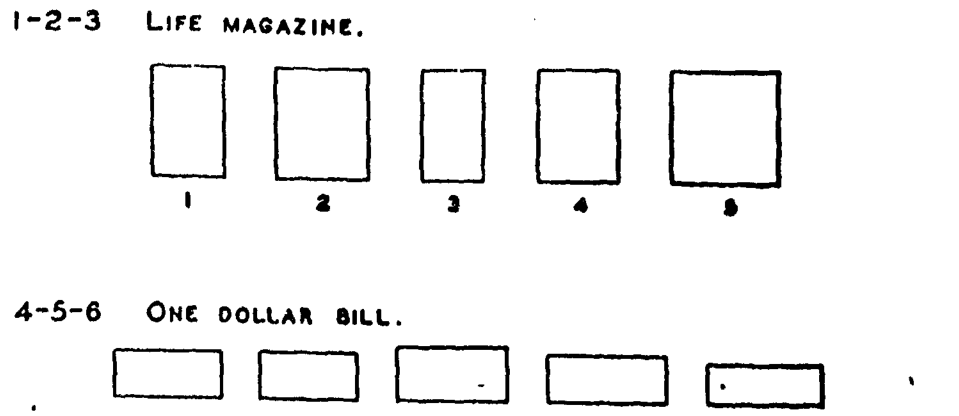


FIGURE 25.4
SAMPLE ITEMS OF PART I OF QUANTITATIVE ESTIMATION, CE440A

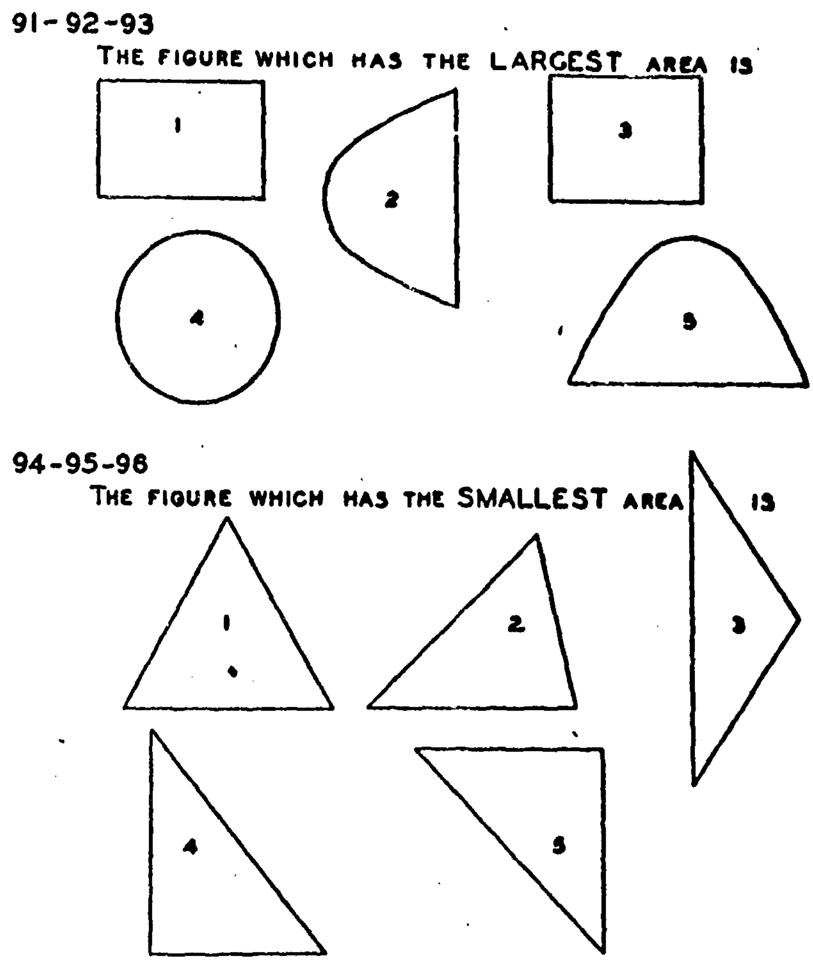


FIGURE 25.5
SAMPLE ITEMS OF PART II OF QUANTITATIVE ESTIMATION
CE440A

703320—47—46

193-194-195.—The weight of two empty coca-cola bottles is:

(1) 28 oz. (2) 32 oz. (3) 24 oz. (4) 20 oz. (5) 36 oz.

202-203-204.—The maximum number of nickels which can be placed flat upon the surface of a dollar bill is:

(1) 21 (2) 32 (3) 28 (4) 18 (5) 24.

184-185-186.—The maximum number of passengers which a standard railroad coach is built to seat is:

(1) 70 (2) 80 (3) 90 (4) 50 (5) 60.

(2) *Administration.*—According to the directions in the booklet, parts I, II, and III were to be timed separately, and examinees were instructed not to proceed to another part until the signal was given. In actual administration, however, these directions were disregarded and an over-all time limit of 35 minutes imposed.

(3) *Scoring.*—The examinee was instructed that if he gave one answer only and it was correct, he would receive 6 points; if he gave two answers and one was correct, he would receive 4 points; and if he gave three answers and one was correct, he would receive 2 points. All incorrect answers were to count zero. For validation, however, the numbers of single, double, and triple responses were the only scores used.

Statistical results.—Preliminary statistics only were obtained concerning proportions making one, two, and three responses and the interrelationships of these data. From the answer sheets of a reportedly large number of preflight individuals tested at Psychological Research Unit No. 1 (classification not identified), 200 were selected at random for analysis. Analysis of these results showed that based on the total number who tried an item, on the average, 62 percent gave one answer, 33 percent gave two, and 5 percent gave three answers. These percentages were computed from data on the first 50 items of the test. Due to the shortness of the testing time, the number of individuals answering items decreased considerably after item 50. For this reason it was determined that the first 50 only should be used in scoring the test for validation.

(1) *Test validity.*—Validation results based on one sample are given in table 25.17.

TABLE 25.17.—*Validity of Quantitative Estimation, CE440A, based upon a sample of pilots in primary training,*[1] *with the graduation-elimination criterion*
[$N_t=556$, $p_s=0.69$]

Score	M_p	M_e	SD_t	r_{bis}	$_c r_{bis}$[2]
Part I—1 answer only	18.81	19.33	7.17	−0.04	−0.01
Part I—2 answers only	9.13	8.73	5.95	.04	.02
Part I—3 answers only	1.65	1.53	2.23	.03	−.01
Part II—1 answer only	12.36	12.72	5.79	−.04	.00
Part II—2 answers only	6.42	6.11	4.94	.04	.01
Part II—3 answers only	.82	.83	1.77	.00	−.03

[1] Tested in June 1944 at Psychological Research Unit No. 1.
[2] Assuming an unrestricted stanine standard deviation of 2.00.

Evaluation.—As evidenced by the validities of various scores for pilot success, this test did not justify the expectations set forth in the hypothesis. It is possible, however, that an instrument of this type might predict success in tasks affected more directly by confidence. On the other hand, examination of parts I and II strongly suggests that these parts measure some perceptual function rather than confidence. Perception and memory may well be predominant also in part III.

Behavior Preference Questionnaire, CE432A [19]

This questionnaire is an attempt to isolate one aspect of personality, namely, degree of self-confidence in social situations.

Description.—This test is of the multiple-choice, personal-inventory type.

(1) *Internal characteristics.*—The questionnaire consists of 40 items, each briefly describing a social-situation problem and presenting four alternate methods of solution. Different degrees of self-confidence presumably are indicated by the various choices. For example:

You are up before a military board for an interview, and the head of the board mispronounces your name. What would you do?

A. Wait until he finishes speaking, then correct him.
B. Correct him at once politely.
C. Say nothing, since it is probably not important.
D. Wait until the end of the interview, then tell him.

You are waiting in line to buy a theater ticket and a man pushes his way in just ahead of you. What would you do?

A. Give him a push out of line.
B. Tell him to go to the end of the line.
C. Comment to those near you in line about the gall of certain people.
D. Do and say nothing.

(2) *Administration.*—The questionnaire is administered as a group test with a time limit of 15 minutes. The test is paced by the administrator by announcement of the time at the conclusion of 5, 10, and 13 minutes, in an attempt to assure completion of a maximum number of items within the time limits of the test.

The directions specify that the examinee indicate how he would actually handle each of the situations described in the test. He is told that there are no right or wrong answers. If he does not definitely prefer any of the alternatives given him, he is required to select the one which comes closest to describing what he would do.

(3) *Scoring.*—In order to derive an a priori score of self-confidence, the alternatives in each item were scaled by several "expert" raters in terms of the degree of self-confidence revealed. Nearly all of the 40 items have 4 alternatives, which were ranked by raters with a relatively high degree of consistency. The test is scored for only the first 30 of the 40 items, since many of the aviation students were unable to complete the test in the time allowed. The least confident answer to each item, as rated by judges, is scored 1 point, while the most confident answer is scored 4, and an intermediate answer either 2 or 3. Thus the

[19] Developed at Psychological Research Unit No. 1. Chief contributor: Lt. Llewellyn N. Wiley.

possible range of scores is from a minimum of 30 to a maximum of 120.

Statistical results.—Data are available for approximately 600 aviation students tested at Psychological Research Unit No. 1 during the period May 31 to June 25, 1943.

(1) *Distribution statistics.*—The range of scores is from 40 to 95 with a median of 69.

(2) *Reliability coefficient.*—The odd-even reliability coefficient obtained with a sample of 205 was 0.22, corrected. Because of the very low reliability of the test, no further analysis of individual scores was made; nor were clinical predictions or ratings of confidence made on the basis of total scores.

(3) *Item validity.*—An item-validation study was made, based on pilots in elementary training. Tetrachoric correlations and levels of significance were computed. The results showed that even the most discriminating response had a tetrachoric *r* significant only at the 2 percent level. In general, the results approximate what might be expected on a purely chance basis. Inspection of the most discriminating items suggested no adequate rationale for their significance.

Evaluation.—On the basis of a very low reliability (0.22 corrected) and purely chance item-validity data, it would seem that this behavior or preference questionnaire is of little promise in predicting air-crew success or success in any other type of endeavor.

A study of response frequencies reveals that for most of the items a large percentage of the examinees selected one alternate in preference to the remaining three, probably because of the strong social approbation connected with that choice. This failure of the items to yield good distributions of responses is ascribed to faults in the wording of the alternates, which left too obvious differences in terms of social desirability or undesirability.

Evaluation of Measures of Confidence

The evidence presented regarding the validity of measures employed in this area is almost entirely inconclusive. This trait, it it be general and consistent in individuals, has proved extremely difficult to quantify. In this respect the trait resembles other traits of temperament, many of which have thus far evaded measurement.

In the light of the results of these tests, it appears that extensive exploration should be made in an effort to find some reliable medium or media for measuring confidence and to determine the amount of communality that exists among tests designed to measure it. The self-ratings of performance on apparatus tests, which probably yielded the purest measures of confidence used, exhibited only moderate correlations with primary pilot graduation-elimination. It may be, of course, that self-confidence is not significantly correlated with pilot success. It seems more likely, however, that a reliable measure of confidence has not yet been discovered.

MEASURES OF SOCIAL INTELLIGENCE AND LEADERSHIP

Analysis of the jobs of fighter and bomber pilots was made to determine the basis or bases upon which a more valid method of assignment could be devised. This analysis indicated that the responsibility of the bomber pilot for other personnel of the crew and the interaction of personalities resulting from the closeness of contacts among crew members demand that the bomber pilot possess high leadership ability. The fighter pilot, on the other hand, has little contact with others while in action, during which he experiences his greatest stress. His social relationships appear, therefore, to be much less significant to the task than are those of the bomber pilot. These findings suggested that some measure of social aptness and leadership ability would assist in singling out the pilots likely to be successful as bomber pilots.

Social Manipulation Inventory, CE443A [29]

It appeared that most of the requirements for such a measure would be met by a social-intelligence test in which problem situations are described and the examinee is required to indicate the best solution. Solutions of the problem situations should involve some understanding of human motivation and of individual differences, with emphasis upon such techniques as the use of praise and blame, delegation of authority, detection and removal of frictions within groups, and the like. If the alternative solutions appear equally plausible and socially acceptable, a good indication should be obtained of what the examinee would do in similar real-life situations.

Description.—This is a purely verbal test similar to the usual judgment test. It is not an inventory of the questionnaire type.

(1) *Internal characteristics.*—The inventory consists of 50 items, each of which depicts a problem situation which might confront an officer or other person having authority over others. In each item there are five alternative courses of action presented. These alternatives were selected as the most appropriate from a list of responses given by unclassified aviation students in free-response interviews. Typical items follow. The responses preceded by an asterisk received a +1 score; all other responses were scored −1.

You are a supervisor of an office force of 10 people. One member is habitually late. You would:

A. Make an example of him by discharging him.

B. Bawl him out in front of the whole group.

*C. Call him in and try to find out the reason for the tardiness.

D. Call a meeting of the office force to explain that everyone owes it to the company to be on time.

E. Call him in privately for a lecture on the importance of being on time.

[29] Developed at Psychological Research Unit No. 3. Chief contributors: S/Sgt. Benjamin Fruchter, Lt. Joseph R. Harsh, Capt. John I. Lacey, Lois G. Wright.

When minor punishment was necessary for your crew, you would:

*A. Hand out the punishment yourself and bring no public attention to it.
B. Have the crew decide the punishment for its members.
C. Turn the matter over to Wing Headquarters to handle.
D. Use sarcasm instead of material punishment.
E. Establish fixed punishments for the usual infractions so that punishment would be automatic.

(2) *Administration.*—The directions to the test emphasize the fact that there are no right answers, but that each person is to respond exactly as he thinks he would under the circumstances described. Although not all possible courses of action are listed, the examinee is required to respond to every item, even if he finds difficulty in deciding which alternative is best.

(3) *Scoring.*—Owing to the nature of the material covered in this inventory, the key was of necessity determined subjectively. After preparation of the form, 12 aviation psychologists and psychological assistants were asked to indicate their judgment as to which alternatives were appropriate and showed the best type of leadership ability. In key A, 36 of the 50 questions were scored. In 7 of these, 2 alternatives were scored as desirable or +1, making a total of 43 desirable responses in the test. The criterion for scoring a response was that 90 percent or more of the judges agree independently as to the desirable and undesirable responses. Almost complete agreement on a large proportion of the items led to the decision not to score the remaining 14 items, pending item-analysis evidence concerning their correlation with the total score on the 36. The formula $R-W/4+20$ was used in scoring with key A, R indicating the number of responses receiving positive weight and W, the number receiving negative weight.

Analysis of the results of the first administration revealed that 10 of the original 36 items had low internal-consistency phi values with total score. A new key (B) was therefore made, scoring 26 of the 36 items scored in key A. The original papers were rescored on this key for rights only.

Item analysis of the original sample scored with key B indicated that some of the 14 originally unscored items were highly correlated with the score obtained with key B. Seven of these items were, therefore, added, and key C was made. The formula $R-W/4+20$ was used with key C.

Statistical results.—Data are limited to distribution constants, estimates of reliability, and a few correlations for samples of pilots in classes 44H and 44I, who were tested in basic training by personnel of Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 25.18. The distribution curves are approximately symmetrical and normal.

TABLE 25.18.—*Distribution constants for Social Manipulation Inventory, CE443A, based upon two groups of classified pilots*

Key	N	M	SD	Scoring formula
B	750	10.2	3.2	Rights only
C	750	18.0	4.6	R−W/4+20

(2) *Internal consistency.*—Analysis of responses based on various keys yielded the internal-consistency data given in table 25.19. The data were all based on the highest 27 percent and lowest 27 percent of groups of 750 classified pilots.

TABLE 25.19.—*Internal-consistency data for Social Manipulation Inventory, CE443A*

Key	Number of scored items	Sample	Mϕ	SDϕ	Range of ϕ	
					Low	High
A	50	I	0.20	0.09	0.02	0.37
A	36	I	.23	.07	.10	.37
B	26	I	.31	.05	.14	.41
C	33	I	.28	.06	.12	.39
C	33	II	.27	.06	.09	.42

(3) *Reliability coefficient.*—Two samples yielded the estimates of reliability given in table 25.20.

TABLE 25.20.—*Estimated reliability coefficients (odd-even) for Social Manipulation Inventory, CE443A*

Group	N	Key	r'_{tt}	r_{tt}
Classified pilots	750	B	0.25	0.40
Classified pilots	750	C	.34	.51

Evaluation.—Although this inventory was not validated, certain pertinent information was obtained in the form of correlations with other measures. The correlation, corrected for attenuation of both variables, with a measure of Reading Comprehension was 0.32, based on an N of 551. The same sample yielded a correlation of 0.50 (corrected for attenuation) with the composite navigator aptitude score. A sample of 556 cases yielded a correlation of 0.15 (corrected for attenuation) with the composite pilot aptitude score.

This evidence seemed to indicate a substantial positive relationship between academic intelligence and the characteristics measured by this inventory. If the hypothesis upon which the measure was based be true, the evidence might indicate that those pilots who have aptitudes most like those of navigators make better bomber pilots. It is possible, on the other hand, that the evidence means that the test measures verbal and other intellectual abilities, but that it does not necessarily imply corresponding intellectual content in the task of the successful bomber pilot.

Its reliability is so low that it would probably be useful only in conjunction with other tests in a battery.

Pilot Behavior Blank, CE444A [21]

As a result of the investigation of pilot specialization, referred to in the report on the Social Manipulation Inventory, it appeared wise to explore the problem of the character or types of leadership. The Lewinian division of leadership into three types—laissez-faire, authoritarian, and democratic—was considered to be a good basis upon which to begin the construction of an instrument for determining the presence and quality of leadership ability. Presumably, such an instrument should be useful in assembling crews of similar tastes with respect to social interaction. A further assumption was that those showing preference for democratic rather than laissez-faire or authoritarian types of leadership should be more successful as bomber pilots where considerable social interaction takes place. By the same token, the type of social interaction preferred by the fighter pilot would be rather unimportant, since he has no crew of his own and exerts little authority over others.

In the light of these assumptions, a preference blank was prepared in which a conscious effort was made to introduce in approximately equal numbers laissez-faire, authoritarian, and democratic solutions to leadership situations without identifying them for the examinee or prejudicing his choice. After a large number of items had been prepared, 10 aviation psychologists and psychological assistants keyed all the choices, indicating which of the three types of leadership each choice indicated (L=laissez-faire, A=authoritarian, D=democratic).

Description.—This blank consists of 90 two-alternative items. The alternatives describe pilots with characteristically different ways of handling situations involving leadership, authority, and the like. The examinee is required to indicate which pilot he prefers. The following samples are typical:

19. A. The pilot who lets the crew members make their own arrangements for quarters, mess, and entertainment.
    B. The pilot who talks about the crew's good points with others.

36. A. The pilot who gives many instructions.
    B. The pilot who while he works more energetically than the rest of the crew, doesn't expect them to work as hard as he.

86. A. The pilot who is so engrossed in his own duties that he hasn't the time to try to understand the difficulties of others.
    B. The pilot who gives his crew exact information for doing a job or carrying out an order.

(1) *Administration.*—Directions on the front page of the blank are intended primarily to prepare the examinee for the material to follow.

[21] Developed at Psychological Research Unit No. 3. Chief contributors: S/Sgt. Benjamin Fruchter, Capt. John I. Lacey, Pfc. Joseph E. Resnner.

Great stress is laid upon the need for knowing what crew members are like, so that congenial individuals can be placed together. "Clash of personalities" is cited as a major cause of incompatibility among members of a group. Examinees are directed to consider the pilots described as always equally capable of flying. The examinee is therefore required to select one of the two alternatives purely on the basis of his own personal preference.

(2) *Scoring.*—The original plan, to score items according to the three Lewinian categories, was discarded in favor of a system more in line with the purpose of the Blank, which was to select men better fitted for assignment as bomber pilots. To accomplish this, the aviation psychologists and psychological assistants examined the alternatives of all items carefully and indicated independently which alternative for each more closely described the good crew commander. Almost unanimous agreement was reached in many of these evaluations. As a result of this study, the choices were keyed for bomber pilot vs. nonbomber pilot, the bomber choices being keyed plus 1 and others minus 1, for 40 selected items out of the total of 90. Sufficient agreement could not be achieved on the remaining 50 items to warrant their being scored. It is interesting to note that the choices selected as good indicators for bomber pilots were in almost every case those designated as "democratic" in the first survey. This fact tends to be justified by the findings of Lewin in regard to the success of different types of leadership. The score is the algebraic sum of the weights of the scored responses.

Statistical results.—Statistics on this test are confined to internally derived data, based upon pilots in classes 44H and 44I, tested during basic training by personnel of Psychological Research Unit No. 3.

(1) *Distribution statistics.*—A sample of 750 classified pilots yielded a mean score of 24.3 and a standard deviation of 9.7, using key A.

(2) *Internal consistency.*—Analysis of responses of several sample groups yielded the internal-consistency data given in table 25.21.

TABLE 25.21.—*Internal-consistency data for items of Pilot Behavior Blank, CE444A, based upon samples of classified pilots*

Sample	N	Key	Number of items scored	$M\phi$	$SD\phi$	Range of ϕ	
						Low	High
I	750	A	31	0.28	0.11	−0.01	0.46
I	750	B	40	.32	.09	.18	.53
II	750	B	40	.31	.09	.16	.51

Key A was the first subjectively-derived key. After analysis of results of the application of this key, it appeared that some choices should be dropped and others added to secure the most internally consistent test. Revision of the key was made in the light of this analysis, and the original sample rescored (key B). A second sample scored with key B yielded similar results.

(3) *Reliability coefficient.* Reliability of this instrument was estimated by the split-half method. The 40 scored items (key B) were divided into 2 groups of 20 each. The items in the two groups were equated for content (laissez-faire, authoritarian, or democratic choices) and for "difficulty." An estimated reliability coefficient of 0.68, corrected for length, was obtained. This figure is based on a sample of 747 classified pilots.

(4) *Difficulty.*—Since there are no right answers to the items in this blank, a difficulty level, strictly speaking, cannot be obtained. The mean proportion of preferred (bomber pilot or democratic) responses is meaningful, however. This figure for the 40 items in key B is 0.58, the range being from 0.23 to 0.87. The mean has not been corrected for chance, since guessing would not be expected to enter a test of this sort where no answers are correct.

Evaluation.—Although validation of the appropriateness of specialization assignment from the standpoint of temperament is very difficult, if not impossible, to accomplish, the psychological values involved indicate that this should be one of the most useful instruments for the selection of persons to exercise authority. It is certain that the underlying democratic principle which is common to the positively scored choices is psychologically sound. One possible weakness of this instrument is that individuals may answer according to known, socially acceptable standards, rather than according to their real bent. Attempt was made to eliminate this type of bias, but it was obviously impossible to make all choices appear equally desi[illegible]. The mean proportion of desirable responses (0.58), however, indicates that this bias was largely eliminated. Evidence that the blank does not involve verbal variance lies in its correlation with Reading Comprehension. A sample of 535 pilots yielded a correlation of only 0.10, corrected for attenuation in both variables. It appears, then, that the blank might well show positive correlation with measures of leadership whenever satisfactory criteria are found.

Evaluation of Measures of Social Intelligence in Leadership

Although statistical proof of the usefulness of the tests described in this section is lacking, certain considerations suggest that the approach employed is one of the most promising in the field of temperament. The trait described here as social intelligence is probably one of the most important determinants of success in personal relations. If this trait can be measured, the results will have far broader significance than for merely air-crew selection or pilot specialization.

Results of investigation in this area revealed or further emphasized difficulties that hinder the construction of reliable measures. In common with many other temperament tests, these instruments are somewhat susceptible to dishonest manipulation. Rapport is therefore extremely important to the achievement of validity. Another danger is that examinees will answer according to considered intellectual judgments rather than according to emotional reactions such as would come into play in face-to-face situations.

EVALUATION OF TESTS OF SPECIFIC TRAITS OF TEMPERAMENT

The heterogeneity of the traits evaluated by the instruments described in this chapter makes it impossible to apply any one evaluative descriptive statement to them. In general, the degrees of success or failure achieved have been noted following the various test or area discussions. Probably the outstanding results reported in this chapter are those obtained in the study of the carefulness tests. The character of the tasks involved and the identity of most of the factors found seem to indicate that the tests resemble aptitude tests more than they do temperament tests. In spite of this fact, however, these findings very strongly suggest that the factor analysis technique should be applied in the study of temperament as well as intellectual measurements.

The experience gained through the development, statistical treatment, and results of tests here reported, has suggested more strongly than ever the necessity of achieving or adhering to certain additional standards in the construction of tests in the temperament area. These standards are notable in this connection, chiefly because they have already become practically axiomatic in the areas of sensation, perception, and intellect.

Probably many requirements or standards could be listed, but four appear to be especially pertinent in this connection. First is the necessity of maintaining a high level of objectivity. It may be argued that complete objectivity cannot be attained in the measurement of traits of temperament. Although this may be true in part, it must certainly be agreed that the maximum attainable objectivity is desirable.

A second requirement is that of reliability. Under this heading might be listed the desirability of having relatively homogeneous material. It appears that many measures in the area of temperament have covered such a wide variety of traits or functions that they constitute a reliable measure of none.

A third requirement is also associated with reliability. It appears necessary to eliminate the element of social acceptability by making alternatives subject to equal or nearly equal social approbation. This goal is obviously difficult to attain when the characteristics being investigated are frequently associated with antisocial behavior. Valid measurement of traits of temperament cannot be achieved, however, by means of instruments that are subject to gross intellectual manipulation.

A fourth requirement has to do with the traits and characteristics selected for study. There should be a logically sound rationale for both the study of the trait and the method to be used. Faulty premises, illogical methods, and irrelevant evidence do not lead to positive results.

The four suggestions listed obviously do not cover all the necessary rules to be observed in the construction of temperament tests. They do cover weaknesses which have been particularly noticeable in the tests reported in this chapter. It is felt that as these requirements are met in test construction, instruments of greater usefulness will result.

CHAPTER TWENTY-SIX

Measures of Motivation[1]

INTRODUCTION

Importance of Measures of Motivation

All individuals concerned with classification of aviation students conceded that it is important to place men in the type of training in which they have the most interest and motivation, or, at least, *not* to assign them to training in which they have little or no interest. Evidence is available that indicates the importance of motivation in both training and combat. This evidence is given in part in chapter 1 and is reviewed in chapter 22.

An Over-all View of Motivation Test Development

In developing measures of motivation, two approaches were taken: (1) a self-assessment by the student through his statement of preferences and (2) a more objective approach through tests of attitudes and interests.

Preference statements.—In the light of the evidence concerning the importance of motivation, arrangements were made to permit students to express their degrees of interest in and preferences for the different air-crew positions. Especially in the beginning of the classification program, and to some extent as late as May 1945, student preferences were used as a guide to classification. At times, quotas seriously interfered with the assignment of men to the air-crew positions they preferred. Unofficial letters and field trips to various flying schools confirmed the expectation that this procedure would cause a considerable lowering of student morale. When the training program was at its peak, therefore, and many candidates were admitted to training, expressed preferences were followed as much as possible in assignment. Besides the limitation of quotas, air-crew aptitude scores became another factor interfering with this policy, particularly as qualifying standards in terms of aptitudes rose. By June 1945 the qualifying scores (stanines) were so high that only a small proportion of the candidates could be assigned to training, and preference statements were no longer obtained.

The appropriate instrument for obtaining the student's statement of preference was not as easy to write as might be supposed. The preference blank went through several revisions, as will be related in the following pages.

The first preference blank, introduced early in 1942, included a list

[1] Written by Sgt. David Grossman.

of eleven air and ground duties. The student was required to rank these in order of his preferences. This form was revised very shortly, however, because of the preponderance of high preferences for air-crew positions. On the revised blank, the student was asked to rank only the three air-crew positions. A blank space provided for his indicating any nonair-crew preference he might have.

The examinee's first preference was given considerable weight in recommending men for bombardier, navigator, and pilot training, when his aptitude scores were relatively uniform. For example: an aviation student indicated his preferences as pilot first, navigator second, and bombardier third. His stanines were 6 for bombardier, 8 for navigator, and 7 for pilot, which qualified him for all three air-crew positions. Because the student's preferences were known, it was possible to recommend him for his first preference; namely, pilot.

The preference waiver.—An additional device known as the preference waiver was also included in this later form of the preference blank. Primarily, the preference waiver was introduced because further information concerning motivation was needed. Over 85 percent of the examinees chose pilot training as their first preference. Since such a large proportion of examinees could not be assigned to pilot training, some means had to be provided for the student to express his willingness to be classified according to the test results rather than by his preferences.

The strength-of-interest scale.—The limitations of the ranking of preferences soon became apparent, and research studies were instituted to devise a better technique for measuring the student's interest. A different type of preference blank, known as the strength-of-interest scale, was constructed, which allowed the examinee to express his strength of interest for each category on a graphic rating scale, with descriptive categories ranging from "little or no interest" to "exceptionally strong interest," as shown in figure 26.1.

FIGURE 26.1
SAMPLE STRENGTH-OF-INTEREST SCALE USED IN THE PREFERENCE BLANK

Both the preference waiver and strength-of-interest scale indicated which students needed to be interviewed prior to classification for a type of training other than their first preference. If a student's stanines were bombardier 6, navigator 8, and pilot 3, his strengths of interest, bombardier 3, navigator 3, and pilot 9, and he had indicated that he would not willingly accept assignment to training other than that of his first preference, he would be interviewed before he was recommended for navigator training.

At the same time that preference blanks were undergoing improvements, efforts were being made to assess motivation more objectively in terms which could be included in the composite aptitude scores. It was never felt that subjective expression of interest was accurate enough, no matter how well scaled, to be included in the stanines.

Tests of attitudes and interests.—Experience with various types of attitude and interest tests will be related in this chapter. They include: Satisfaction, CE409A, B, C, and D; Aviation Preference Check List (no code); Inventory of Experiences, Interests, and Attitudes, CE612-AX2; Specialization Preference Inventory, CE610A; Specialization Interest Inventory, CE 609A; Social Concepts, CE 512A; Survey of Personal Attitudes, CE 508A; Inventory of Attitudes, CE518A; Conduct of the War Test, CE520A; and Home Front Attitude Inventory, CE446A.

PREFERENCE BLANKS

Aviation Cadet Training Preference Blank, CE501E[3]

The preference blank differs from most classification instruments in that it is not a test. Form E of the blank will be described in much detail, since it was used over the longest period of time, and since it is the result of the accumulated experience gained in using previous forms, CE501A and D, and CE 509A.

Description.—In the first part of the blank the examinee is asked to state his degree of interest in each type of training by encircling the number which represents that degree of interest. A sample scale is shown in figure 26.1.

There are three scales, one for each of the three types of air-crew training. If the examinee has a stronger interest in any other type of training, he can name it and express this interest on a fourth scale.

There are many reasons for using a graphic rating scale rather than a ranking of preferences as in the A and D forms of the blank. In the first place, the linear scale allows for ties. Second, it permits an aviation student to show unequal differences in interest between his first and second, and his second and third preferences. Ranking of preferences implies equal distances. Third, it presents a satisfactory way in which an examinee can show lack of interest for one or more types of air-crew training. Fourth, it allows the examinee to consider each air-crew position separately, on its own merits, instead of in relation to the other two positions. The examinee is forced to think much more carefully, since he is required to express himself in more specific terms. Finally, this method enables the examinee to express his first preference as somewhat below maximum strength. A student who gives ranked preferences of bombardier *2*, navigator *1*, and pilot *3*, may not have "exceptionally strong interest" for navigator, as ranking might imply.

[3] Developed at Psychological Research Unit No. 1. Chief contributor: Maj. Frederick Wickert.

On the other hand, graphic rating scales have some disadvantages. Because the linear scale did allow for ties, interviews were often required before classification recommendations were made. The most serious disadvantage is that one may be misled into assuming that the points on the scale have the same meanings for all examinees.

The second part of the preference blank is called the preference waiver. The examinee must check one of the following statements:

1. I want to be assigned to the kind of air-crew training for which I show the greatest ability on the tests.
2. I want to be assigned to the kind of air-crew training for which I show the greatest ability on the test only if my ability for that kind of training is much greater than for any other kind.
3. I want to be assigned to the kind of air-crew training in which I am most interested unless the tests show that I should probably fail in that kind of training.
4. I want to be assigned to the kind of air-crew training in which I am most interested even if the tests show that I should probably fail in that kind of training.

These statements are worded simply and in nontechnical terms. It was found that many students experienced serious difficulty in comprehending the statements in the earlier Form D. The wording was too academic and was written from the standpoint of a psychologist rather than from that of an aviation student. During interviews, students expressed inability to understand such concepts as "distinctly higher," "aptitude," or "prediction."

(1) *Administration.*—The preference blank was administered at the beginning of the first session of group testing, and it required approximately 10 minutes. No time limit was set. The following are excerpts from the directions to the preference blank:

It is important that every cadet be assigned to a kind of air-crew training in which he can succeed. Two factors that determine how well a cadet will succeed are:

A. His scores on the classification tests, which measure how much ability he has for bombardier, navigator, and pilot training.
B. How strong an interest he has in each type of training. Evidence shows that a cadet is more successful in a type of training in which he is intensely interested.

In stating your interests, you should consider these matters carefully:

A. How much you know about the duties of each member of the air crew.
B. Whether your own ability, education, and training in your own judgment fit you for one kind of work rather than another.
C. How much you desire each type of training, and how *willing* you are to work hard to succeed in it.

Statistical results.—Many research studies were instituted in connection with the preference blank. As a result, many data exist. A representative sampling is submitted here.

(1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this blank are given in table 26.1. Distributions of students according to preference waiver are shown in table 26.2.

TABLE 26.1.—*Means and standard deviations of ratings on the strength-of-interest scale administered at the time of classification testing*

Group	N	M_B[1]	M_N	M_P	SD_B	SD_N	SD_P
Pilots in primary training[2]	707	5.6	5.1	8.5	1.9	2.0	1.1
Do[3]	11,423	5.5	5.9	8.6	1.9	2.1	.9
Navigators in advanced training[4]	1,953	5.1	7.2	7.9	1.9	1.9	1.6

[1] M_B stands for mean strength-of-interest rating in bombardier training, etc.
[2] In class 44B. Tested at Psychological Research Unit No. 3.
[3] In class 44C. Tested at Psychological Research Units Nos. 1, 2, and 3.
[4] In classes 43–12 to 43–15 inclusive. Tested at Psychological Research Units Nos. 1, 2, and 3.

TABLE 26.2.—*Percentages of aviation students selecting each type of preference waiver administered at the time of classification testing*

Group	N	Type of waiver[1]			
		1	2	3	4
Pilots in primary training[2]	701	38.5	12.5	44.9	3.8
Do[3]	9,551	29.3	10.2	57.1	3.4
Navigators in advanced training[4]	1,554	34.0	14.4	46.4	5.2

[1] For types see page 000.
[2] In class 44B. Tested at Psychological Research Unit No. 3.
[3] In class 44C. Tested at Psychological Research Units Nos. 1, 2, and 3.
[4] In classes 43–12 to 43–15 inclusive. Tested at Psychological Research Units Nos. 1, 2, and 3.

(2) *Validity of strength-of-interest ratings.*—In tables 26.3 to 26.9 are presented the results of numerous studies of the relation between graduation-elimination from various types of air-crew training and the variables of strength of interest and of first preference (i. e., air-crew position receiving highest strength-of-interest rating).

TABLE 26.3.—*Relation of first preference to graduation-elimination for samples of bombardier, navigator, and pilot trainees*

Group[1]	N	p_e	Obtained chi-square	df[2]	Values required for significance	
					5 percent	1 percent
Bombardiers—12-week course[3] .	1,706	0.88	5.79	4	9.49	13.28
Bombardiers—18-week course[4] ..	455	.84	3.31	3	7.82	11.34
Navigators in advanced training[5]	1,953	.79	22.80	3	7.82	11.34
Pilots in primary training[6]	11,423	.84	97.55	3	7.82	11.34
Pilots in basic training[7]	6,702	.87	1.47	3	7.82	11.34

[1] All samples consist of examinees from the three psychological research units.
[2] df=degrees of freedom.
[3] In classes 43–15 to 44–1 inclusive.
[4] In classes 43–14 to 43–18 inclusive. The 18-week course, unlike the 12-week course, includes some training in navigation.
[5] In classes 43–12 to 43–15 inclusive.
[6] In class 44C.
[7] In class 44I.

TABLE 26.4.—*Relation of first preference to graduation-elimination of pilot trainees from primary training* ($N_t=11,423$,[1] $p_e=0.84$)

First preference	P_t[2]	P_{te}[3]	r_{bis}
Bombardier	2.9	1.2	(4)
Navigator	7.6	5.8	−0.16
Pilot	83.2	71.7	.12

[1] In class 44C. Tested at Psychological Research Units Nos. 1, 2, and 3.
[2] Percentage preferring type of training.
[3] Percentage preferring type of training and graduating.
[4] Split too extreme for computation of r_{bis}.

703320—47—47

Table 26.5.—*Percentages eliminated by pilot stanine and first preference, based on 5,501 pilots in primary training*[1]

Pilot stanine	First preference		
	Bombardier: N=98	Navigator: N=254	Pilots: N=5,149
7-9	28.6	23.9	13.2
4-6	43.8	54.0	29.6
1-3	70.4	70.8	49.7
Total	50.0	47.6	28.4
Mean pilot stanine	4.45	5.51	5.34

[1] In class 43F. Tested at the three Psychological Research Units.

Table 26.6.—*Relation of strength-of-interest rating to graduation-elimination of pilot trainees from primary training* ($N_t=11,423$,[1] $p_g=0.84$)

Interest category	M_g	M_e	SD_t	r_{bis}[2]
Bombardier	5.47	5.72	1.93	−0.07
Navigator	5.88	6.13	2.10	−.07
Pilot	8.66	8.50	.91	.10

[1] In class 44C. Tested at the three Psychological Research Units.
[2] A biserial *r* of approximately 0.03 is required for significance at the 5 percent level, and of 0.04 at the 1 percent level.

Table 26.7.—*Relation of strength-of-interest to graduation-elimination of navigator trainees from advanced training* ($N_t=1,953$,[1] $P_g=0.79$)

Interest category	M_g	M_e	SD_t	r_{bis}[2]
Bombardier	5.13	4.97	1.92	0.05
Navigator	7.34	6.84	1.94	.15
Pilot	7.89	7.99	1.61	−.03

[1] In classes 43-12 to 43-15 inclusive. Tested at the three Psychological Research Units.
[2] A biserial *r* of approximately 0.06 is required for significance at the 5 percent level and of 0.08 at the 1 percent level.

Table 26.8.—*Relation of strength-of-interest to graduation-elimination of 1,706 bombardier trainees from 12-week course* ($N_t=1,706$,[1] $P_g=0.88$)

Interest category	M_g	M_e	SD_t	r_{bis}[2]
Bombardier	6.94	6.83	2.03	0.02
Navigator	5.69	5.49	2.34	.05
Pilot	7.68	7.70	1.81	−.01

[1] In classes 43-15 to 44-1 inclusive. Tested at the three Psychological Research Units.
[2] A biserial *r* of approximately 0.08 is required for significance at the 5 percent level and of approximately 0.10 at the 1 percent level.

Table 26.9.—*Relation of strength-of-interest rating to graduation-elimination of bombardier trainees from the 18-week course*[1] ($N_t=513$,[2] $P_g=0.86$)

Interest category	M_g	M_e	SD_t	r_{bis}[3]
Bombardier	7.00	6.87	2.04	0.03
Navigator	5.25	4.96	2.44	.07
Pilot	7.60	7.85	1.86	−.07

[1] The 18-week course included some navigation training.
[2] In classes 43-14 to 43-18 inclusive. Tested at the three Psychological Research Units.
[3] A biserial *r* of approximately 0.13 is required for significance at the 5 percent level and of approximately 0.18 at the 1 percent level.

Tables 26.3 and 26.4 show that there is a slight but significant relationship between graduation from primary pilot training and first preference. This relation, however, does not hold for basic training (table 26.3). Similarly, table 26.6 shows a slight but significant relationship between primary pilot training and strength of interest. These results are confirmed by the data as shown in table 26.5. At each stanine level, those whose first preference is for bombardier or navigator training have a considerably higher elimination rate from primary pilot training than those whose first preference is for pilot training.

First preference and strength of interest for navigation have significant correlations with success in advanced navigation training (see tables 26.3 and 26.7).

Bombardier first preference and strength of interest show no relation to the criterion (see tables 26.3, 26.8, and 26.9).

(3) *Validity of the preference waiver.*—Tables 26.10 to 26.12 give the results of studies to determine the relation between the preference waiver and graduation-elimination from various types of air-crew training.

TABLE 26.10.—*Relation of preference waiver to graduation-elimination for samples of bombardier, navigator, and pilot trainees*[1]

Group	N	p_s	Obtained chi-square	df[2]	Values required for significance	
					5 percent	1 percent
Bombardiers—12-week course[3]	1,706	0.88	1.65	3	7.82	11.34
Bombardiers—18-week course[4]	455	.84	10.35	3	7.82	11.34
Navigators in advanced training[5]	1,953	.79	6.00	3	7.82	11.34
Pilots in primary training[6]	11,423	.84	4.44	3	7.82	11.34
Pilots in basic training[7]	6,702	.87	9.49	3	7.82	11.34

[1] All samples consist of examinees of the three Psychological Research Units.
[2] Degrees of freedom.
[3] In classes 43–15 to 44–1 inclusive.
[4] In classes 43–14 to 43–18 inclusive.
[5] In classes 43–12 to 43–15 inclusive.
[6] In class 44C.
[7] In class 43I.

TABLE 26.11.—*Relation of preference waiver to graduation-elimination of pilot trainees from primary training (N_t=11,423,[1] P_s=0.84)*

Preference waiver[2]	P_t[3]	P_{tp}[4]	r_{bis}
1	29.3	24.2	−0.02
2	10.2	8.6	.06
3	57.1	48.0	−.01
4	3.4	2.8	([5])

[1] In class 44C. Tested at the three Psychological Research Units.
[2] For categories see page 724.
[3] Percentage of total group selecting waiver.
[4] Percentage of total group selecting waiver and graduating.
[5] Split too uneven for computation of r_{bis}.

The preference waiver does not seem to bear much relation to graduation-elimination (see tables 26.10 and 26.11). Two significant relations are found between the 5 percent and 1 percent levels; one in a sample of pilots in basic training and the other in a sample of bombardiers taking the 18-week course. None of the other groups shows relationships significantly

TABLE 26.12.—*Percentages eliminated by pilot stanine and preference waiver based on 7,826 pilots in primary training*[1]

Pilot stanine	Preference waiver[2]			
	1	2	3	4
7–9	19.4	15.4	11.1	5.6
4–6	36.4	30.1	27.5	24.9
1–3	60.2	49.4	51.3	46.0
Total	36.9	28.6	27.1	23.7
Mean pilot stanine	5.11	5.46	5.32	5.34

[1] In class 43F. Tested at the three Psychological Research Units.
[2] For categories see page —.

different from zero. It can be seen readily in table 26.12, however, that those students who check waiver 4 (I want to be assigned to the kind of air-crew training in which I am most interested EVEN IF the tests show that I should probably fail in that kind of training) are less likely to be eliminated at each stanine level.

(4) *Relationship of strength of interest and first preference to the stanines.*—Studies were made to investigate the relation between the stanines and preferences and waivers. It also seemed desirable to compute the differences in mean stanine scores for students expressing different first choices. The data are shown in tables 26.13, 26.14, and 26.15.

TABLE 26.13.—*Correlations of strength of interest with stanines of 700 unclassified aviation students*[1, 2]

Stanine	Interest category		
	Bombardier	Navigator	Pilot[3]
Bombardier	−0.02	0.19	0.07
Navigator	−.14	.29	.08
Pilot	−.12	.09	.05

[1] Tested at Psychological Research Unit No. 3 in December 1942 and January 1943.
[2] A product-moment r of approximately 0.07 is required for significance at the 5 percent level and of 0.10 at the 1 percent level.
[3] Correlations in this column are biserials contrasting strength-of-interest rating of 9 with all others. The p_e is unknown.

TABLE 26.14.—*Mean stanines by first preferences of 678 unclassified aviation students*[1]

First preference	Stanine		
	Bombardier	Navigator	Pilot
Bombardier	4.81	4.38	5.53
Navigator	5.49	5.30	5.95
Pilot	5.13	4.78	5.81

[1] Tested at Psychological Research Unit No. 3 in December 1942 and January 1943.

TABLE 26.15.—*Critical ratios of differences between mean stanines of students preferring one type of training and mean stanine of students preferring all other types, based upon 678 unclassified aviation students*

First preference	Stanine		
	Bombardier	Navigator	Pilot
Bombardier	−3.94	−5.50	−2.35
Navigator	3.81	6.57	1.41
Pilot	−2.77	−2.80	−0.66

In general, there is but slight relation between strength of interest and the stanine. It will be seen in table 26.13, however, that the strength of interest for navigator training has significant positive correlations with all three stanines, especially with the navigator stanine. First preference for bombardier training has very low but significant negative correlations with navigator and pilot stanines. Tables 26.14 and 26.15 show that students whose first preference is for navigator training have higher average stanines in all three specialties, those for bombardier and navigator aptitude being significantly different from the general means of all other students. Students whose first preference is for bombardier training have significantly lower mean stanines in all three specialties. (See also table 26.5.)

(5) *Relation of stanine validity to first preference and preference waiver.*—An analysis was made to test the hypothesis that the pilot stanine would have higher validity for those whose first preference is for pilot training. The results are reported in table 26.16. Similar results were computed for the preference waiver and are presented in table 26.17.

TABLE 26.16.—*Biserial correlations of pilot stanine[1] with graduation-elimination from primary training when pilot trainees[2] are grouped by first preference*

First preference	N	Pilot stanine validity
Bombardier	328	0.43
Navigator	864	.49
Pilot	9,512	.45
Navigator or bombardier	1,192	.46

[1] Derived from the classification battery of December 1942.
[2] In class 43F. Tested at Psychological Research Unit No. 3.

TABLE 26.17.—*Biserial correlations of the pilot stanine[1] with graduation-elimination from primary training when pilot trainees[2] are grouped by preference waiver*

Preference waiver[3]	N	Pilot stanine validity
1	2,251	0.39
2	2,432	.37
3	2,852	[4].43
4	291	[5].50

[1] Derived from the classification battery of December 1942.
[2] In class 43F. Tested at Psychological Research Unit No. 3.
[3] For categories see page —.
[4] Significantly different from 0.37, with a critical ratio of 2.6.
[5] Significantly different from 0.39 and 0.37, with critical ratios of 2.2 and 2.6 respectively.

It can be seen in table 26.16 that the validity of the pilot stanine is *not* higher for those whose first preference is pilot training. On the contrary, the validity of the pilot stanine is slightly higher for pilots whose first preference is for navigator training. None of the differences is statistically significant, the highest critical ratio being 1.2. In table 26.17, it may be seen that the preference waiver has a rather significant relationship to the validity of the pilot stanine. The predictive value of the stanine is highest for the men who chose preference waiver 4.

(6) *Relationship between strength of interest and classification-test scores.*—To ascertain the relation between tests in the August 1942 battery and the intensity of motivation of aviation students, intercorrelations were computed. This study is important, because certain tests may be regarded on an a priori basis as indirect measures of interest. Some, indeed, were designed as such (e. g., Technical Vocabulary, CE 505C). The results are shown in table 26.18.

TABLE 26.18.—*Correlations of strength of interest with tests of the August 1942 classification battery, based on 707 pilot trainees*[1, 2]

Test and code number	Interest category		
	Pilot[3]	Bombardier	Navigator
Technical Vocabulary (Pilot), CE505C	0.02	−0.10	0.04
Technical Vocabulary (Bomb.), CE505C	−.04	−.02	.11
Technical Vocabulary (Nav.), CE505C	−.07	−.18	.22
Speed of Identification, CP610A	−.02	.00	.07
Mathematics, CI702E	−.12	−.20	.34
Numerical Approximations, CI706A	−.13	−.01	.21
Reading Comprehension, AC10D	−.08	−.08	.20
Mechanical Comprehension, AC10D	−.16	−.18	.03
Table Reading, CP621A	.06	.02	.14
Numerical Operations, CI702B	−.07	.03	.13
Spatial Orientation I, CP501B	−.07	−.03	.06
Spatial Orientation II, CP503B	−.03	−.12	.03
Arithmetic Reasoning, CI206B	.07	−.04	.19
Dial Reading, CP622A	.01	−.10	.13
Complex Coordination, CM701A	−.02	−.06	.04
Steadiness, CM103A	−.06	−.02	.00
Finger Dexterity, CM116A	.04	.04	.06
Discrimination Reaction Time, CP611D	.01	.01	.15

[1] In class 44B. No previously eliminated students are included. Tested at Psychological Research Unit No. 3.

[2] A product-moment r of approximately 0.07 is requir-d for significance at the 5 percent level and of approximately 0.10 at the 1 percent level.

[3] Biserial correlations. A biserial r of approximately 0.10 is required for significance at the 5 percent level and of approximately 0.13 at the 1 percent level.

A few important findings should be noted from table 26.18. Strength of interest for pilot training shows significantly negative relationships with three tests. These tests are Mechanical Comprehension (with a correlation of −0.16), Mathematics (−0.12), and Numerical Approximations (−0.13). Six tests show significantly negative correlations with bombardier interest. Strength of interest for navigator training has significant positive correlations with eleven tests. Of these, Technical Vocabulary (navigator), Mathematics, Numerical Approximation, Reading Comprehension, and Arithmetic Reasoning are the tests that have proved themselves most valid for success in navigation training.

The Technical Vocabulary and Information Test (CE505C) was designed as an interest test, with three interest scores, each for one of the air-crew specialties. Correlations of first preference for a specialty with the score for that specialty are negative and not significantly different from zero except for navigator preference and navigator score (see table 26.19).

If the expressed specialized preferences are good criteria of specialized interest, only one score on this test (navigator) proves to be valid for the purpose intended. There is other evidence (see p. 817) that the pilot

TABLE 26.19.—*Relation of first preference to Technical Vocabulary and Information, CE505C, for unclassified aviation students*[1]

Score	First preference		
	Pilot	Bombardier	Navigator
Pilot	[2]−0.08	....	....
Bombardier	....	[3]−0.05	....
Navigator	....	....	[2]0.53

[1] Biserial correlations, the dichotomy being those whose first preference is for the indicated air-crew position *v.* those whose first preference is for all other positions.

[2] $N_t=527$, $P_q=0.91$. A biserial of 0.15 is required for significance at the 5 percent level and of 0.20 at the 1 percent level. Group tested in July 1943 at Psychological Research Unit No. 3.

[3] $N_t=530$, $P_q=0.95$. A biserial of 0.18 is required for significance at the 5 percent level and 0.23 at the 1 percent level. Group tested in July 1943 at Psychological Research Unit No. 3.

score *does* measure pilot interest to some extent, which leads us to suspect expressed pilot preference as a criterion of pilot interest.

(7) *Preferences in relation to pilot specialization.*—This study was part of a larger project designed to study the problem of differentiating among aptitudes for various types of advanced specialized pilot training. Assignment to specialized training to an appreciable degree was a matter of preference.

The means and standard deviations for the strengths of preferences in the three air-crew specialties are given in table 26.20. In table 26.21 critical ratios are presented for the differences between mean strengths of interest of students in different types of advanced training. Two critical ratios indicate differences significant at or beyond the 5 percent level. These are the critical ratios of the difference (1) in navigator training interest for those assigned to fighter training and those assigned to heavy bomber training, and (2) in pilot training interest for those assigned to fighter or medium-bomber training. The first ratio indicates that pilots who are assigned to training on heavy bombers express more interest in navigation than those who become fighter pilots. The difference between medium and heavy bombardment assignees, though not significant, is in the same direction. Assignees to fighter-pilot training expressed a more intense desire for pilot training, in general, than those students given medium-bomber training. The same trend is indicated between fighter and heavy-bomber training.

TABLE 26.20.—*Mean strength of interest for trainees*[1] *in fighter, medium bomber, and heavy bomber planes*

Interest category	Fighter training[2]		Medium bomber training[3]		Heavy bomber training[4]	
	M	SD	M	SD	M	SD
Pilot	8.73	0.71	8.55	1.07	8.63	0.92
Bombardier	5.14	1.81	5.39	1.94	5.16	1.92
Navigator	4.53	2.00	4.56	2.16	4.92	2.15

[1] In class 43J. Tested at Psychological Research Unit No. 3.

[2] N=565.

[3] N=196.

[4] N=607.

TABLE 26.21.— *Critical ratios[1] of differences between mean interest-strength in types of advanced pilot training*

Category	Fighter v. medium bomber training	Fighter v. heavy bomber training	Medium v. heavy bomber training
Bombardier	−1.54	−0.17	1.42
Navigator	−0.18	−2.36	−1.47
Pilot	2.12	1.82	−0.90

[1] Positive critical ratios indicate that those assigned to training in lighter planes express greater strength of interest in the given interest category.

Variations of Preference Blanks.—There were five other preference blanks, two of which—Aviation Cadet Training Preference Blanks, CE501A and D—were used in the classification battery before December 1942.

(1) Aviation Cadet Training Preference Blank, CE501A and D.—These earlier blanks were developed to allow aviation students to indicate their training interest, as well as to obtain a measure of intensity of motivation.

In form A, the student is asked to rank his preferences from 1 to 11 for the types of duties listed. This list includes: armament officer, bombardier, communications officer, engineering officer, gunner, mechanic, meteorologist, navigator, photographer, pilot, and radio operator.

In form D the three types of air-crew duties are listed, and the students are instructed to write 1 opposite their first choice, 2 opposite their second choice, and 3 opposite their third. They are required to mark all three, and not to give any two the same rank. It was with this form that the preference-waiver section was introduced. Following are the statements presented for the student's use:

1. I would prefer to be classified for the type of duty for which I am found to have most aptitude, even though it is not the same as the first preference given above.
2. I would prefer to be classified for the type of duty for which I am found to have most aptitude, only if my aptitude for this type of duty is at least two points higher on the 9-point aptitude scale than for the duty for which I expressed first preference.
3. I would prefer to be classified for the type of duty for which I am found to have most aptitude, only if my aptitude for my first preference indicates that I am likely to be eliminated from that type of training. (A score of 3 or below.)

I would prefer to be assigned to the type of training for which I indicated first preference above without regard to my aptitude scores.

The difficulties and limitations of the ranking-preference technique and of the wording of the preference-waiver section were discussed on pages 723f.

(2) Aircrew Preference Rating Scale, CE503A and B.[6]—The circumstance that prompted the development of these scales was the dissatisfaction with the ranking method used in Training Preference Blank, CE501D.

[6] Developed at Psychological Research Unit No. 1. Chief contributor: Maj. Frederick Wickert.

In form A, students are asked to compare their preferences for each type of air-crew training paired with every other type.

The items for form B were taken from items scaled for the Navy. The examinee's task is to check in the column headed "Pilot" those statements that describe his feelings about pilot training. The procedure then is repeated under the columns for "Navigator" and "Bombardier."

Form A was not given to a sufficiently large sample to make possible the computation of any statistics. This form lacked the advantage of giving an independent intensity-of-preference score for each of the three types of air-crew duty. The principal value of this work was its indication of what not to do in later forms.

Tabulations of the data for form B indicated that practically the same items were checked for pilot, navigator, and bombardier. Differentiation of interests for the three types of training, therefore, was not great enough to be useful. The results did not correspond with informal impressions of the degree to which preferences for the three types of training were differentiated in the minds of the students. Since the score obtained for each of the three types of training appeared to be so meaningless, no validation was attempted.

(3) Aviation Cadet Preference Scale, CE509A.—This is a graphic rating scale, and it is the last variation of the blank designed to secure preferences for training. It consists of a horizontal line divided into 11 spaces by dots placed at equal intervals along the line. One end of the line is labeled "Dislike intensely," the center "Indifferent," and the other end "Like intensely." No other descriptive comments are used. The student is asked to draw a short vertical line through a dot and write B, N, and P (for bombardier, navigator, and pilot) on the line that represents how he feels about each type of training.

This scale, modified in accordance with suggestions from the Office of the Air Surgeon and from Psychological Research Unit No. 3, resulted in a new form called the Aviation Cadet Training Preference Blank, CE501E, which was accepted for use in the classification battery and was described above.

Training Preference Blank, CE513A

This blank was developed in order to measure varying degrees of interest, not in types of training, but in types of airplanes, combat and noncombat. It is comparable to the blank filled out in basic training to aid in the assignment of graduates to various types of advanced training. The hypothesis was that such a blank would be valid for the graduation-elimination criterion in primary and other phases of training.

Description.—Nine-point scales like those appearing in the Aviation Cadet Training Preference Blank, CE501E, are utilized in this blank (see fig. 26.1).

(1) *Internal characteristics.*—The first section consists of seven scales, listing various types of planes: trainers, transports, fighters (single and twin-engine), and bombers (both medium and heavy). The examinee is instructed to encircle the number which corresponds to his interest, ranging from 1 ("little or no interest") to 9 ("exceptionally strong interest").

In section two, the student is presented with a list of the seven planes, which he is to rank according to his preferences from 1 to 7.

In section three, another graphic scale is presented, and the student is asked to circle the number which represents how disappointed he would be if he had to learn to fly a type of plane not among his first three choices.

(2) *Administration.*—The approximate administration time for the blank is 10 minutes, but no time limit is set.

Statistical results.—(1) *Distribution statistics.* Typical examples of distribution statistics obtained on this blank are given in table 26.22.

TABLE 26.22.—*Distribution constants for Training Preference Blank, CE513A, based on 1,130 classified pilots entering primary training*[1]

Type of plane	M	SD
Trainer	4.55	2.22
Transport	4.85	1.97
Twin-engine fighter	6.97	2.01
Medium bomber (B-25)	6.27	1.74
Medium bomber (B-26)	5.91	1.88
Single-engine fighter	7.11	2.11
Four-engine bomber	6.27	2.14
Disappointment scale	4.38	1.86

[1] In class 44F. Tested at Psychological Research Unit No. 3.

(2) *Test validity.*—This blank was administered for validation to classified pilots just prior to entrance into primary training. Validation data for strength of preference (on the 9-point scale for each of the various types of planes) are presented in table 26.23.

TABLE 26.23.—*Relation of strength of preference for different types of planes to graduation-elimination of pilots from primary training*[1] *(N_t=1,130, p_g=0.90)*

Type of plane	M_g	M_e	SD_t	r_{bis}[2]	$_cr_{bis}$[3]
Trainer	3.56	3.40	2.22	0.03	0.03
Transport	3.86	3.76	1.97	.02	.02
Twin-engine fighter	6.02	5.38	2.01	.15	.17
Medium-bomber (B-25)	5.28	5.07	1.74	.06	.05
Medium bomber (B-26)	4.93	4.62	1.88	.08	.07
Single-engine fighter	6.15	5.59	2.11	.13	.11
Four-engine bomber	5.28	5.08	2.14	.05	.03
Disappointment scale	3.40	3.05	1.86	.09	.10

[1] In class 44F. Tested at Psychological Research Unit No. 3.
[2] A biserial correlation of approximately 0.10 is required for significance at the 5 percent level and of approximately 0.13 at the 1 percent level.
[3] Assuming an unrestricted stanine standard deviation of 2.00.

Evaluation.—From the table of correlations, it can be seen that expressed interest for only two types of planes, single-engine and twin-engine fighters, shows a significant relation to graduation-elimination from primary training. All coefficients, however, are positive. As for all interest ratings, these correlations are attenuated by individual constant errors, but the extent of attenuation is an open question.

Evaluation of Air-Crew Preference Blanks

It was originally hoped that the strength-of-interest scale and the preference waiver would indicate intensity of motivation. Interview results, however, showed that many other factors enter into the student's decision as to which descriptive category to check and which statement to mark. The following are typical:

(1) Many students seem to distort their ratings of strength of interest in order to influence the classification board. An example would be a student giving a very high preference for one position (such as "9" for pilot), and placing the other two very low ("1" for bombardier and "2" for navigator) on the scale.

(2) Many students, perhaps in fear of being grounded, signify their preferences by circling "9" for all categories.

(3) Students tend to circle numbers that have words under them, i. e., the odd numbers (see fig. 26.1).

(4) Being unable to understand what the statement means, the student may mark a statement because he is influenced by irrelevant factors, such as the feeling-tone imparted by the wording of the statement.

(5) Marking statements in the preference-waiver section often indicates whether the student has faith in the tests or not, thereby measuring his attitude toward psychological tests as well as (or rather than) his motivation.

(6) Many students seem to mark the waivers "1" or "2" because of a sense of duty or patriotism in doing what the Army wants them to do.

(7) Many students mark waivers "1" or "2" because of a feeling that the classifying officer will be prejudiced against them if they mark statements "3" or "4."

(8) Some students seem to feel that the four waiver statements constitute a disguised personality test of some sort. The students feel that they will be considered indecisive or lacking in a knowledge of their own desires if they mark statement "1," and that they may be considered stubborn if they mark statement "4."

There are some grounds for the belief that the preference waiver should be given at the end of the testing sessions rather than at the beginning. After the students have had some experience with the tests, they feel better qualified to state whether they desire to be classified according to the test results.

There are a few general conclusions and group tendencies that can be noted in an examination of the statistical data. In brief, they are:

(1) Those students who prefer pilot training are most likely to succeed in graduating from primary training (but this does not apply to basic training).

(2) Those whose first choice is bombardier have the lowest average stanines for all three air-crew positions.

(3) Those whose first choice is navigation training have the highest average stanines for all three air-crew assignments. Conversely, those who put navigator last in their order of preferences have the lowest average stanines.

(4) The validity of the pilot stanine is not higher for those whose first choice is pilot training.

(5) Strength of interest for navigation training has significant positive correlations with classification tests that have high validity with the navigator criterion, while strength of interest for either pilot or bombardier training does not correlate well with any tests.

Because navigator interest is significantly related to the navigator stanine, success in navigator training, and the classification tests, there is reason to believe that students who prefer navigator training have superior insight into their abilities and temperament.

The strong emphasis placed upon the student's preferences in making recommendations for air-crew training is not warranted by the empirical evidence. This is not to say that the information yielded by the preference blank was not valuable to those recommending air-crew classifications. The conclusion is undeniable, however, that self-assessed preferences have very low validity for predicting success in air-crew training.

ATTITUDE AND INTEREST INVENTORIES

Various types of instruments—personal inventories, situation tests, check lists, and the like—were used to obtain expressions of specific likes and aversions. This constitutes an indirect and more objective method, as contrasted with the direct and less objective method of general self-assessment. It was assumed that these specific preferences would form a pattern and serve as a basis of trainee selection.

Satisfaction Test, CE109A [4]

This test was developed on the basis of certain hypotheses concerning interest-trends in the various air-crew specialists. The successful combat pilot was assumed to be extroverted; the navigator was assumed to be characterized by sedentary and scientific interests; and the bombardier was assumed to be characterized by aggressive-destructive tendencies.

Description.—A purely empirical approach to the measurement of air-crew personality was adopted in developing this inventory. A collection was made of verbally described personality-revealing situations by studying "Information Essays" written by students in connection with the development of Technical Vocabulary and Information Test (CE505A)[5] and by studying job analyses containing personality data.

[4] Developed at Psychological Research Unit No. 1. Chief contributors: Tech./Sgt. Robert R. Blake, Capt. Donald E. Super, Staff/Sgt. John L. Wallen.

[5] See chapter 14 for a discussion of this test.

Preliminary research on expressions of likes and dislikes of aviation students involved the preparation of two questionnaires. In Questionnaire I, the examinees were given descriptions of a number of situations in which soldiers frequently find themselves. The examinees were asked to write briefly what they would do in each case. In Questionnaire II the examinees were asked to write down those features of military life—different from civilian life—which they liked and those which they disliked. A statement of the number of months the examinee had been in active service was also obtained. In order to secure complete frankness on both questionnaires, the examinees were told that this information would have no bearing upon their classification, and they were told not to sign their names on their papers.

The data obtained from the second questionnaire furnished material for part I (Aviation Cadet Likes) and part II (Aviation Cadet Dislikes) of the Satisfaction Test, CE409A. It was found that "Likes" result from personal care and privileges (food, regular schedules, uniforms, equipment, etc.); social relations (comradeship, discipline, social uniformity, etc.); and personal values (educational opportunities, chance for advancement, freedom from worry about present or future, exactness, precision, promptness, etc.). "Dislikes" result from physical inconveniences (food and method of serving, lack of lights, mud, crowded or inadequate toilets and washing facilities, etc.); duties or lack of privileges (getting up early, latrine duty, K. P., guard duty, not enough free time, etc.); personal frustrations (loss of individuality, harsh authority, being away from home and friends, etc.); inefficiency (organization waiting in line, lack of knowledge about the future, slowness to get action on matters of personal importance, etc.); and social relations (other cadets considered undesirable, lack of feminine companionship, etc.).*

A few examples of the types of items in part I and II are:

As a cadet, I would get more satisfaction from:

(A) Everybody being on equal terms, or
(B) Getting regular medical care.
(A) Security and freedom from worry about the present, or
(B) The chance for a career.
(A) The chance for a career, or
(B) Good pay.

As a cadet, I would be more irritated by:

(A) Unfair or harsh orders, or
(B) Being away from home and loved ones.
(A) Too much regimentation, or
(B) Lack of recreational facilities.
(A) Lack of personal privacy, or
(B) Lack of feminine companionship.

* It is interesting to note the differences between new men and those longer in the service. It appeared that (1) individual differences in dislikes are largely eliminated with increasing military experiences, so that all men gradually come to dislike very nearly the same things; (2) personal frustrations play a more important part as length of service increases; and (3) the general area of social relationships becomes of increasingly greater importance with time in the service, serving as the main source both of obstacles and of aid to personal adjustment in army life.

The data obtained from Questionnaire I proved very useful in constructing items for part III, "cadet opinions," of the satisfaction test. This part is composed of a number of described situations, each followed by descriptions of five alternative responses. The five alternatives were chosen in accordance with six assumed modes of reacting to a problem situation: rationalization, reaction-formation, compensation, introversion (submission), extraversion (ascendance), and neuroticism.

It was decided that there would be no alternative referring to an integrated manner of behaving, since it seemed probable that it would appear to the examinees to be the obviously right answer. It was expected that the examinee would either (1) consistently choose alternatives of a given kind, revealing behavior which is characteristic of himself, or (2) choose alternatives revealing various modes of reaction so that his score for any one mode of reaction would be moderate. The latter result was thought to identify the integrated person. A typical part III item is:

If a fellow is not genuinely admired by the men in his squadron, he will most probably:

- **A. Feel that the other men are not the kind that he cares to be admired by and make no effort to gain their friendship.**
- **B. Try to be the best soldier in the bunch in order to show them up.**
- **C. Say that someone who is jealous of him is passing rumors about him.**
- **D. Feel glad because he gets more satisfaction from being on his own than from being one of the gang.**
- **E. Feel hurt by their coolness and try to boss them around.**

In the item above, A would be the rationalization response, B the compensation, C the neurotic, D the submissive-introverted reaction, and E the ascendant-extroverted reaction.

Part IV, Preferences, contains paired comparisons of civilian activities (sports and hobbies) and flight assignments assumed to be related to personality traits. Following are examples of items appearing in Part IV.

If given the choice and having equal opportunity and ability, would you rather:

- **(A) Go to a carnival? or**
- **(B) Go to an opera?**
- **(A) Play cards? or**
- **(B) Play football?**
- **(A) Be a sergeant on active flying duty? or**
- **(B) Be a lieutenant with a desk job?**

(1) *Internal characteristics.*—In final form the test consists of 150 items divided among the four parts just described. Part I contains 30 items; part II, 30; part III, 45; and part IV, 45.

(2) *Administration.*—Before each part, a brief introductory statement is made. For parts I, II, and IV the examinee is told to judge each pair independently of the other pairs, since any alternative may occur several times, paired with a different alternative each time. The examinee is instructed that if little preference is felt for either alternative, he is to choose the less objectionable activity.

For part III, the examinee is instructed to choose the alternative which most closely corresponds to his own opinion about each particular situation or activity. If none of the statements exactly expresses the opinion of the examinee, he is instructed to choose the one that is the closest to his opinion. Even if several choices could express the opinion of the examinee, he is instructed to select only the best one.

The total testing time for part I is 10 minutes; for part II, 10 minutes; for part III, 16 minutes; and for part IV, 10 minutes.

The test is paced, with the administrator interrupting the testing to ask for a show of hands of those who had finished part I at the end of 7 minutes, part II at the end of 7 minutes, part III at the end of 12 minutes, and part IV at the end of 7 minutes.

(3) *Scoring.*—The papers were first scored on two a priori keys prepared by the test constructors, one designed to measure atypicality of attitude and one designed to measure morale. Since these keys had no validity, a priori keys were abandoned in favor of empirically derived keys.

Following two item analyses based on random halves of a total group of 1,595 cases, using as the criterion graduation-elimination from primary pilot training, two empirical scoring keys were prepared for cross-validation. All responses that were made by significantly different percentages of graduates and eliminees (significant at the 10 percent level or better) were examined to discover whether the items could be defended psychologically as well as statistically. On these bases, two scoring keys were constructed.

The scoring formula is R−W+40, in which R refers to the positively weighted responses and W to the negatively weighted.

Statistical results.—The data that follow are for examinees tested in October 1942 at Psychological Research Unit No. 1.

(1) *Distribution statistics.*—Typical examples of distribution statistics obtained on this test are given in table 26.24.

TABLE 26.24.—*Distribution constants for Satisfaction Test, CE409A, for pilots in primary training*

N	M	SD
787	40.0	6.3
75[illegible]	46.4	5.7

(2) *Test validity.*—Based on the two a priori keys, one to measure atypicality of attitude and one to measure morale, a sample of 787 classified pilots yielded biserial correlations with graduation-elimination in primary pilot training of −0.12 and −0.01 respectively. Cross-validation data for the two empirically developed keys, however, are much more satisfactory. The data appear in table 26.25.

TABLE 26.25.—*Cross-validation data for Satisfaction Test, CE409A, based on pilots in primary training, using graduation-elimination as the criterion*

Group	Key	N_t	p_g	M_g	M_e	SD_t	r_{bis}
Evens	Odds	797	0.58	40.8	39.0	6.3	0.18
Odds	Evens	798	.58	47.3	45.4	5.7	.21

Evaluation.—The a priori keys showed no validity for primary pilot training. The empirical keys, however, developed on the basis of item-validity coefficients, yielded validities of approximately 0.2 upon cross-validation. The validity remained at about 0.2 in subsequent administrations of two revisions of the test (see below). This validity and the low correlations with the pilot stanine indicate that the test would be expected to add moderately to the validity of the classification battery.

The item-analysis data indicated that the rationale back of the construction of this test, as far as the pilot is concerned, was promising and that there was a need for further item revision and item writing. On the strength of the results, the test was revised and two new forms prepared.

Satisfaction Test, CE409B and C[7]

These two forms are revisions and extensions of CE409A.

Description.—The items for Forms B and C were selected from parts III and IV of Form A of the test on the basis of the item validities revealed in the previously discussed analyses. The criterion for selection was a tetrachoric correlation of 0.08 or more with success in primary pilot training, which is at the 10 percent level of significance or better. In addition to these, new items were constructed according to the principles which seemed to underly the types of items already demonstrated to be valid.

(1) *Internal characteristics.*—Forms B and C consist of 85 paired-comparison items, 25 in part I and 60 in part II. In Form B, the examinee is required to make a choice between the two alternatives. Form C, however, which contains the same items as Form B, permits a third response —"Neither of these." Items in part I are prefixed with different premises, such as the two items below. Asterisks indicate the alternatives with positive pilot validity.

Other things being equal, actual parachute jumping would be:

*A. A thrilling experience.
B. Good because it is so necessary for fliers.
C. Neither of these.[8]

As a cadet, I would expect to get more satisfaction because:

*A. I would like to fly.
B. Fliers are badly needed.
C. Neither of these.[8]

[7] Developed at Psychological Research Unit No. 1. Chief contributors: Tech./Sgt. Robert R. Blake, Capt. Donald E. Super, Staff/Sgt. John L. Wallen.
[8] These alternatives appear only in form C.

All items in part II are prefixed with the same premise, such as:

If given the choice and having equal opportunity and ability, would you rather:

A. Do high level precision bombing?
*B. Do dive bombing?
C. Neither of these.[*]

*A. Practice aerobatics?
B. Practice instrument flying?
C. Neither of these.[*]

(2) *Administration.*—The total testing and administration time for Form B or C is 24 minutes. Parts I and II are not timed separately.

(3) *Scoring.*—The scoring formula for Forms B and C is R−W+20, in which R refers to positively weighted responses and W to negatively weighted responses, as determined by item analyses. The constant is added to eliminate negative scores.

Statistical results. (1) *Distribution statistics.*—Typical examples of distribution statistics obtained for Satisfaction Test, CE409B and C, are given in table 26.26.

TABLE 26.26.—*Distribution constants for Satisfaction Test, CE409B and C based on pilots in primary training*[1]

Form	N	M	SD
B	739	12.0	4.6
B	740	7.2	4.2
C	566	6.2	6.0
C	529	9.6	6.3

[1] In classes 44F to 44J inclusive. Tested at Psychological Research Unit No. 1 in August and September 1943.

(2) *Reliability coefficient.*—The internal consistency of the test is indicated by the Kuder-Richardson (Formula No. 21) coefficients in table 26.27.

TABLE 26.27.—*Estimated reliability coefficients (Kuder-Richardson Formula No. 21) for Satisfaction Test, CE409B and C, based on pilots in primary training*[1]

Form	N	r
B	739	0.71
B	740	.67
C	566	.82
C	529	.86

[1] Same samples as in table 26.26.

(3) *Item validity.*—After dividing the sample of 1,479 answer sheets (the first two samples of table 26.26) into two random halves, the responses to the items in each sample of Form B were correlated with the graduation-elimination criterion from primary pilot training in order to develop an empirical key. The same procedure was followed with the 1,095 answer sheets (the last two samples of table 26.26) of Form C. The distributions of the tetrachoric *r*'s are shown in table 26.28. Items

[*] These alternatives appear only in form C.

703320—47—48

are not included in this distribution when one response was chosen by more than 90 percent or less than 10 percent of the examinees.

TABLE 26.28.—*Tetrachoric r distribution for items in satisfaction test, CE409B and C, based on odds and evens samples of pilots in primary training, using graduation-elimination as the criterion*

Tetrachoric r	Form B[1]		Form C	
	f		f	
	Odds[2]	Evens[3]	Odds[4]	Evens[5]
0.28–0.32	....	....	1	2
0.23–0.27	....	....	1	1
0.18–0.22	2	4	4	6
0.13–0.17	6	7	14	14
0.08–0.12	20	20	19	17
0.03–0.07	27	23	33	34
−0.02–0.02	17	26	31	46
(−0.07)-(−0.03)	27	23	37	23
(−0.12)-(−0.08)	20	20	18	20
(−0.17)-(−0.13)	6	7	11	11
(−0.22)-(−0.18)	2	4	10	6
(−0.27)-(−0.23)			4	2
Total	127	134	163	182

[1] Note that this is the 2-choice form of the test.
[2] $N_t=740$, $P_g=0.77$.
[3] $N_t=739$, $P_g=0.79$.
[4] $N_t=529$, $P_g=0.80$.
[5] $N_t=566$, $P_g=0.79$.

In interpreting these tetrachoric *r*'s, it can be said that for the N of approximately 750 an r_{tet} of 0.07 is significant at the 5 percent level, and an r_{tet} of 0.09 is significant at the 1 percent level of confidence. For the N of approximately 550, an r_{tet} of 0.09 is significant at the 5 percent level, and an r_{tet} of 0.11 is significant at the 1 percent level of confidence. In the odds sample of Form B, 31 questions yield r_{tet}'s, which exceed the 5 percent level of significance and 23 questions which exceed the 1 percent level. In the evens sample of Form B, the corresponding figures are 33 and 24. In the odds sample of Form C 67 r_{tet}'s exceed the 5 percent level of significance and 53 exceed the 1 percent level. In the evens sample of Form C, the corresponding figures are 72 and 58.

(4) *Cross-validation data.*—Cross-validation results using the keys based on these item validations are given in tables 26.29 and 26.30. The same procedure was followed for Form C. The data appear in tables 26.29 and 26.30.

TABLE 26.29.—*Cross-validation data for Satisfaction Test, CE409B, for samples of pilots in primary training, based on graduation-elimination criterion*

Sample	Key based on sample	Scoring formula	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_{c}r_{bis}$[1]
I (odds)	II (evens)	R−W+20	[2]739	0.79	12.68	11.37	4.60	0.18	0.18
II (evens)	I (odds)	R−W+20	[2]740	.77	7.85	6.42	4.24	.20	.20
III	I and II	Rights	[3]1,475	.79	19.01	17.48	3.81	.23	.24
III	I and II	Wrongs	[3]1,475	.79	8.10	9.43	3.75	−.20	−.21

[1] Corrected to an unrestricted stanine standard deviation of 2.00.
[2] Same samples as in table 26.26.
[3] New sample tested at Psychological Research Unit No. 1 in August and September 1943.

TABLE 26.30.—*Cross-validation dc'a for Satisfaction Test, CE409C, for samples of pilots in primary training, based on graduation-elimination cri'erion*

Sample	Key based on sample	Scoring formula	N_t	r_e	M_s	M_u	SD_t	r_{bis}	$_cr_{bis}$[1]
I (odds) ..	II (evens) .	R−W+20	[2]566	.79	7.03	5.17	6.03	0.18	0.19
II (evens)	I (odds) ...	R−W+20	[2]529	.80	10.48	8.60	6.32	.17	.18
III	I and II ...	Rights ...	[3]1,399	.80	18.47	16.72	4.26	.23	.25
III	I and II ...	Wrongs ..	[3]1,399	.80	8.84	10.69	3.98	−.26	−.28

[1] Corrected to an unrestricted stanine standard deviation of 2.00.
[2] Same samples as in table 26.26.
[3] New sample tested at Psychological Research Unit No. 1 in August and September 1943.

The scoring key used for sample III is the combined key developed from the item analysis of the odds and evens sample. Because Forms B and C are for all intents and purposes the same test, it was possible to utilize the four existing item-validation studies in making a final key. This key consists of 38 positively weighted alternatives and 36 negatively weighed alternatives. The scoring formula is R−W+20, in which R refers to positively weighted responses and W the negatively weighted responses.

(5) *Use of satisfaction test for pilot specialization.*—An item analysis was completed for this test, using as the criterion the ratings given to pilots on general pilot ability.[6] Forty-two bomber pilots who were above average in general pilot ability in transition schools and 97 fighter pilots who were above average in single-engine advanced schools were used. In interpreting the phi coefficients, it can be said that for an N of 150, a phi of 0.16 is significant at approximately the 5 percent level, and a phi of 0.21 is significant at the 1 percent level of confidence. In this sample, 11 items exceed the 5 percent level of significance and 7 exceed the 1 percent level. This tends to show that if items that differentiate between bomber and fighter pilots were keyed, a prediction could be made that would help to place pilots properly in their specialties. The distribution of the phi coefficients is shown in table 26.31.

TABLE 26.31.—*Distribution of phi coefficients for Satisfaction Test, CE409B, based on a sample of 148 pilots in transition and advanced training*[1]

Phi[2]	Frequency	Phi[2]	Frequency
0.30–0.34	1	0.10–0.14	16
0.25–0.29	1	0.05–0.09	17
0.20–0.24	5	0.00–0.04	23
0.15–0.19	7	Total	70

[1] In class 44B. Tested while in basic training by personnel of Psychological Research Unit No. 3.
[2] Since the test is a two-choice form, only positive phis are tallied.

Tables 26.32 and 26.33 present means, standard deviations, and critical ratios of the differences in the mean scores on the Satisfaction test for high and low bomber pilots and high and low fighter pilots. For the purpose of obtaining the high and low groups, the Pilot Proficiency

[6] Ratings of "above average," "average," and "below average" were given to pilots by instructors and flight commanders on the student's pilot proficiency cards at each phase of training.

Cards were consulted. The ratings bomber pilots received in general pilot ability in advanced and transitional training were summed, and two groups were formed—those at or above the mean and those below the mean. The same was done for fighter pilot ratings in basic and advanced training. The four groups then were scored with the final empirical key discussed on page 26.46.

TABLE 26.32.—*Means and standard deviations of high and low bomber and fighter pilots*[1] *on Satisfaction Test, CE409B*

Group	Score	N	M	SD
High bombers	Rights	213	19.9	5.2
	Wrongs	213	15.8	4.7
	R−W+20	213	24.1	8.7
Low bombers	Rights	120	22.3	5.7
	Wrongs	180	13.6	4.7
	R−W+20	180	28.7	9.4
High fighters	Rights	144	25.1	5.2
	Wrongs	144	10.6	3.5
	R−W+20	144	34.4	7.3
Low fighters	Rights	164	25.2	4.9
	Wrongs	164	10.9	3.7
	R−W+20	164	34.3	7.3

[1] In class 44B. Tested in basic training by personnel of Psychological Research Unit No. 3.

TABLE 26.33.—*Critical ratios of differences of means of "high" and "low" bomber and fighter pilots on Satisfaction Test, CE409B*

Groups	Score	CR
"High" bomber v. "low" bomber	Rights	[1]4.4
	Wrongs	[1]4.7
	R−W+20	[1]5.0
"High" fighter v. "low" fighter	Rights	.0
	Wrongs	.7
	R−W+20	.1
"High" bomber v. "high" fighter	Rights	[1]9.3
	Wrongs	[1]12.0
	R−W+20	[1]12.1
"Low" bomber v. "low" fighter	Rights	[1]5.0
	Wrongs	[1]5.9
	R−W+20	[1]6.3

[1] Significant at the 1 percent level.

Evaluation.—Forms B and C of the Satisfaction test proved to be very useful testing instruments. Compared with form A, they are shorter in length, take half the time, and are easier to administer. On three different samples, the test maintained a moderately high validity coefficient with the pilot criterion. There is also reason to believe that this inventory measures some factor not adequately measured by the classification battery. Owing to its low correlation with the stanine and its moderate validity for pilot criterion, this test would be a valuable addition to the pilot-selection battery.

It is interesting to note the nature of the items that show consistent validity for pilot selection. Examination of the keyed items suggests definite clustering around the picture of the eager fighter pilot, who likes to fly for the sake of flying, for the excitement, and for personal adventure. This interest and attitude pattern appears to be related to success in training, at least at the primary level.

Form B seems promising as a vehicle to determine in which specialty (bomber *v.* fighter) a pilot should go. As in the aviation preference check list, the statistical analysis indicates that there are personality traits which differentiate successful bomber pilots from successful fighter pilots.

There are 12 responses (one-third of all keyed items) in form B that are weighted positively for the primary pilot-validity key, positively for the fighter-pilot key, and negatively for the bomber-pilot key. There are five responses which are positively weighted both for students in primary pilot training and for bomber pilots in transitional training. The other 19 items weighted for primary pilot are weighted zero for pilot-specialization criteria. These data tend to show that primary-training motivation is more akin to fighter-pilot motivation than to bomber-pilot motivation, and this is corroborated by the data in tables 26.32 and 26.33.

As has previously been explained, form C differs from form B only in that it has a third alternative—"Neither of these." On the basis of the Kuder-Richardson formula, form C appears to provide a somewhat more reliable test (0.8 as against 0.7 for form B). Because this slightly higher reliability is not accompanied by a correspondingly higher validity, it is felt that the evidence does not warrant using the third response in constructing a new form.

Variation of the test (1) Satisfaction Test, CE409D.[a]—Form D is the culmination of all the work done on the Satisfaction test and the Aviation Preference Check List. Because the two tests were quite similar, it was possible to choose the most valid items from a combined total of 235 items. The 60 items of highest validity were selected. The correlation of these items with the pass-fail criterion in primary pilot training is indicated by an absolute mean phi of 0.11 and a range from 0.07 to 0.19. Twenty additional items were added to these 60 for padding to confuse the examinee as to the purpose of the test and so to keep him from rationalizing which are the right answers.

Form D is divided into two parts of 40 items each. The items of the two parts are matched for validity, content, and percentage answering the item.

The total testing and administration time for form D is 16 minutes, with parts I and II timed separately.

For the form D key, the 60 alternatives with previous positive validity were given weights of +1, the 60 alternatives with previous negative validity were given weights of −1, and 0 weights were given to the 20 padding items. The scoring formula is R−W+20, the constant being added to eliminate negative scores.

Form D is shorter and takes less administration and testing time than the previous forms. Because of the matched parts, it is expected to give a better estimate of reliability. Correspondingly higher validity is also

[a] Developed at Psychological Research Unit No. 2.

expected because of the inclusion only of items with high validity coefficients.

No statistical data are available.

Aviation Preference Check List (no code)

This check list is a revision of the Navy Aviation Preference Check List. It was hoped that differences would be found in the likes and dislikes of successful and unsuccessful pilots. If such an instrument proved to be successful for selecting pilots, similar instruments might also assess the motivation for bombardier and navigator training.

Description.—The preference list is composed of descriptions of activities in the fields of sports, hobbies, and military flight assignments. The activities are paired, and the examinee is required to choose the activity he prefers. The two alternatives represent different types of participation in the same field of activity. The alternatives are worded to eliminate bias as much as possible, since analysis has shown that certain adjectives or adverbs tend to influence responses to items.

(1) *Internal characteristics.*—In the final form of the test there are 150 items. A typical item is: "(A) Fly with others in the ship, (B) Fly solo."

(2) *Administration.*—The examinee is instructed to answer items quickly according to his first impression, avoiding deliberation. He is required to answer all questions. No time limit is set, but 20 minutes are sufficient time for almost everyone to finish. The test is paced by the administrator. After 6 minutes, the examinees are told that they should be on item 50. After 12 minutes, they are told that they should be on item 100.

(3) *Scoring.*—The scoring formula is $R-W+40$, in which R represents responses with positive weights and W responses with negative weights.

Statistical results.—(1) *Item validity.* After dividing a sample of 1,459 answer sheets completed by examinees at Psychological Research Unit No. 3 into two random halves, the responses to the items in the sample were correlated with a modified graduation-elimination criterion, with students rated "below average" in any phase of training[a] placed with the eliminees. Graduates had passed at all levels of training, and they include both bomber and fighter pilots. The distributions of the phi coefficients are shown in table 26.34. Items are not included if one response is chosen by more than 90 percent or less than 10 percent of the examinees.

In interpreting these phi coefficients, it can be said that for an N of 730, a phi of 0.07 is significant at approximately the 5 percent level, and a phi of 0.09 is significant at the 1 percent level of confidence. In the odds sample, 23 items exceed the 5 percent level of significance, of which 16 exceed the 1 percent level. In the evens sample, the correspond-

TABLE 26.34.—*Distribution of phi coefficients for Aviation Preference Check List based on pilots at all levels of training, using a modified graduation-elimination criterion*

Phi[1]	f: Odd[2]	f: Even[3]
0.15–0.19	2	0
0.10–0.14	10	4
0.05–0.09	46	42
0.00–0.04	73	79
Total	121	125

[1] Since the test is a two-choice form, only positive phis are tallied.
[2] N_t=729, p_q=0.77. In classes 44F and 44G.
[3] N_t=730, p_q=0.77. In classes 44F and 44G.

ing figures are 21 and 6. There were 39 items keyed in the odds sample and 27 items keyed in the evens sample.[11] Of these, only eight items were keyed the same on both samples. Six items were keyed with opposite signs, and 45 items were keyed zero on one sample and positive or negative on the other. The discrepancy in the two keys is probably due to a preponderance of bomber pilots in the evens sample. That there is such a preponderance is indicated by the fact that the same items that proved valid, in the study reported below on pilot specialization, for discriminating between fighter and bomber pilots, have the most significant positive phis in the evens sample only.

(2) *Use of Aviation Preference Check List for pilot specialization.*—An item analysis was completed for this check list, using as the criterion General Pilot Ability[8] ratings. Fifty-two bomber pilots, who were rated "above average" in general pilot ability in transition schools, and 97 fighter pilots, who were rated "above average" in single-engine advanced schools, were used. The distributions of the phi coefficients are shown in table 26.35.

TABLE 26.35.—*Distribution of phi coefficients for Aviation Preference Check List based on a sample of 52 bomber pilots in transition training and 97 fighter pilots in advanced training*[1]

Phi[2]	Frequency	Phi[2]	Frequency
0.29–0.33	3	0.09–0.13	40
0.24–0.28	17	0.04–0.08	35
0.19–0.23	14	0.00–0.03	41
0.14–0.18	17	Total	167

[1] In class 44B. Tested in basic training by personnel of Psychological Research Unit No. 3.
[2] Since the test is a two-choice form, only positive phis are tallied.

In interpreting these phi coefficients, it can be said that for an N of 150, a phi of 0.15 is significant at the 5 percent level, and a phi of 0.21 is significant at the 1 percent level of confidence. In this sample, 44 phis exceed the 5 percent level of significance and 27 exceed the 1 percent level. This would tend to show that if items that differentiate between bomber and fighter pilots were keyed, a prediction could be made that would help to place pilots properly in their specialties.

[11] Some items were keyed that did not quite reach the 5 percent level of significance, if other data showed that the items could reasonably be expected to be valid.

Tables 26.36 and 26.37 present means, standard deviations, and critical ratios of the differences in mean scores on the Aviation Preference Check List for "high" and "low" bomber pilots and "high" and "low" fighter pilots. For the purpose of obtaining the "high" and "low" groups, the Pilot Proficiency Cards were consulted. The ratings bomber pilots received in General Pilot Ability in advanced and transitional training were summed, and two groups were formed—those at or above the mean and those below the mean. The same was done for fighter-pilot ratings in basic and advanced training. The four groups then were scored with the empirical key, valid for the primary graduation-elimination criterion.

TABLE 26.36.—*Means and standard deviations of high and low bomber and fighter pilots on Aviation Preference Check List*[1]

	Score	N	M	SD
High bombers	Rights	213	18.9	4.8
	Wrongs	213	21.2	4.9
	R−W+40	213	38.0	9.6
Low bombers	Rights	180	22.6	6.0
	Wrongs	180	17.9	6.1
	R−W+40	180	45.0	11.9
High fighters	Rights	144	26.4	4.5
	Wrongs	144	13.6	4.2
	R−W+40	144	53.0	8.7
Low fighters	Rights	164	26.3	4.6
	Wrongs	164	14.1	4.4
	R−W+40	164	52.0	9.3

[1] In class 44B. Tested in basic training by personnel of Psychological Research Unit No. 3.

TABLE 26.37.—*Critical ratios of differences of means of high and low bomber and fighter pilots on Aviation Preference Check List*

Groups	Score	CR
High bomber v. low bomber	Rights	[1]6.6
	Wrongs	[1]6.0
	R−W+40	[1]6.3
High fighter v. low fighter	Rights	.2
	Wrongs	1.0
	R−W+40	.9
High bomber v. high fighter	Rights	[1]15.0
	Wrongs	[1]15.7
	R−W+40	[1]15.3
Low bomber v. low fighter	Rights	[1]6.4
	Wrongs	[1]6.6
	R−W+40	[1]6.2

[1] Significant at or beyond the 1 percent level.

Evaluation.—The Aviation Preference Check List seems promising in connection with pilot specialization. Statistical analysis of the items suggests that there are personality traits which differentiate successful bomber pilots from successful fighter pilots. A definite preference pattern appears, indicating that the fighter pilot likes to fly for the sake of flying, for the excitement, for the personal adventure, and that he has a devil-may-care attitude. Bomber pilots appear to be more conscientious, methodical, thorough, and [illegible]ring, and must be willing to accept responsibility for men and e[illegible].

Some examples of items that differentiate enough to support these descriptions of the two types of pilots are given below. Asterisks indicate responses positively weighted for bomber pilots:

*A. Fly with others in the ship.
B. Fly solo.
A. Work involving few details.
*B. Work involving many details.
A. Aerobatics.
*B. Instrument flying.
*A. Read a book.
B. Read a magazine.
A. Repair a motor.
*B. Tell others how to.
A. Strafe hostile infantry.
*B. Bomb hostile fort.
*A. Fight from formation.
B. Fight individual dogfights.
*A. Ground School Instructor.
B. Physical Training Instructor.

Inventory of Experiences, Interests, and Attitudes, CE612AX2 [11]

This is another inventory designed to assess personal background and preferences.

Description.—The inventory consists of a number of questions concerning the examinee's past experiences and activities, his interests and preferences, his feelings and attitudes. The majority of the items are of the multiple-choice type. Some of the activities, however, are paired, and the examinee is required to choose one of the two activities of each pair. These paired items are quite similar to the items in the Aviation Preference Check List. In the third section of the inventory different jobs are described, and the examinee is asked to rate the job on a scale ranging from "very much more attractive" to "very much less attractive." The last ten questions consist of a list of 10 airplanes. The examinee is asked to express on a five-point scale how he would feel about being assigned to training on each of the planes. Some of these last questions were used as the criterion for the rest of the test.

(1) *Internal characteristics.*—There are 150 items, divided into 4 parts: part I contains 82 items; part II, 40 items; part III, 20 items; and part IV, 10 items. Typical items of part I are given below. The plus and minus signs indicate weights of +1 and −1 derived from item-validation procedures to be described below.

How often do you write your parent or parents?
0 A. Parents not living.
−B. Almost every day.
−C. About twice a week.
+D. About once a week.
+E. Less frequently than once a week.

Do you spend a good deal of time planning what you wish to do after the war?
+A. Yes.
−B. No.

[11] Developed at headquarters AAF Training Command and Psychological Research Unit No. 2. Chief contributors: Maj. R. L. Thorndike and Maj. S. Rains Wallace.

Two typical items of part II follow. The examinee is required to choose one of the alternatives.

−A. Do stunt flying.
+B. Do straight flying.

−A. Acrobatics.
+B. Instrument flight.

In part III, the examinee rates jobs on a five-point scale: (A) very much more attractive, (B) more attractive, (C) neither more nor less attractive, (D) less attractive, and (E) very much less attractive.

Two typical job-descriptions in this part are:

In this job the pay consists mostly of commissions on sales. (B is keyed minus and D plus.)

This job calls for a high degree of skill in athletics. A is keyed minus and B plus.

In part IV, the examinee indicates his job satisfaction, using a five-point scale: (A) would be extremely well satisfied with this assignment; (B) would be well satisfied with this assignment; (C) would be moderately well satisfied with this assignment; (D) would prefer a different assignment; and (E) would very much prefer a different assignment.

Using this scale, he expresses his satisfaction with an assignment to training in the following planes: B–17, B–24, B–29, B–25, B–26, A–20, P–38, P–49, P–47, and P–51.

(2) *Administration.*—While there are no time limits for this test, experience has shown that all should be finished at the end of 35 minutes. It has been found advisable to pace individuals at the end of 15 minutes by saying, "You should be on item No. 70."

(3) *Scoring.*—To secure a criterion for the development of an empirical key, answer sheets of bomber-pilots in transitional training were scored on the questions concerning the B–17, the B–24, and the B–29. This provided a possible range of scores from 3 through 15, since an A response was scored as five points, B as four, C as three, D as two, and E as one. Answer sheets of fighter-pilots in advanced training were scored on the questions concerning the P–38, the P–40, the P–47, and the P–51, thus providing a possible range of scores from 4 to 20. These were the criterion scores which were taken to represent degrees of satisfaction.

Approximately 1,800 bomber-pilots and 1,600 fighter-pilots had taken the inventory. They were tested in transitional and advanced training, respectively, in both Western and Central Flying Training Commands, by personnel of Psychological Research Unit No. 2. Each set of papers was divided into two random halves, and item analyses were accomplished. On the basis of these item analyses, two keys were constructed in a manner described below.

The scoring formula is R−W+20, in which R represents responses with positive weights and W responses with negative weights.

Statistical results.—As described above, this test was administered to pilots in advanced and transitional training, after they had had experience in flying specialized planes. The test scores (criterion scores described above) for the inventory were reduced to a 9-point scale, with an attempt to obtain a mean of 5 and a standard deviation of 2.

Item analysis.—The papers were split at random into four groups: bomber-pilot odds and evens samples, and fighter-pilot odds and evens samples. Using the highest and lowest 25 percent of the subsamples, on the basis of the criterion scores, 6 item analyses were accomplished. The six item analyses contrasted high *v.* low fighter-pilot groups, high *v.* low bomber-pilot groups, and high bomber *v.* high fighter-pilot groups, for the odds and evens samples separately.

Two criteria were used for the inclusion of a response in a key: the response had to differentiate the high-bomber-pilot group from the high fighter-pilot group with a phi of 0.15 or better (beyond the 1 percent level of confidence), and it had to differentiate either the high-fighter-pilot group from the low-fighter-pilot group or the high-bomber-pilot group from the low-bomber-pilot group with a phi of 0.10 or better (beyond the 5 percent level of confidence). All phis were computed from percentages based on the total number answering a question, rather than the total number taking the test.

The test was keyed for bomber-pilot satisfaction. In the final odds keys, 31 responses were keyed positively, and 32 responses were keyed negatively. In the final evens key, 41 responses were keyed positively, and 48 were keyed negatively. No responses were keyed negatively in one sample that were also keyed positively in the other, or vice versa. The number of significant responses at the 1 and 5 percent levels are presented in table 26.38.

TABLE 26.38.—*Number of significant responses for Inventory of Experiences, Interests, and Attitudes, CE612AX2, based on pilots tested in advanced and transition training*

Analysis	Sample	N	Number of significant responses	
			5 percent	1 percent
High bomber pilots v. low bomber pilots	Odds	[1]450	136	95
Do	Evens	[1]450	144	99
High fighter pilots v. low fighter pilots	Odds	[2]400	82	56
Do	Evens	[2]400	106	56
High bomber pilots v. high fighter pilots	Odds	[2]400	115	72
Do	Evens	[2]400	119	85

[1] A phi of 0.09 is required for significance at the 5 percent level, and of 0.12 at the 1 percent level.

[2] A phi of 0.10 is required for significance at the 5 percent lev·'. and of 0.13 at the 1 percent level.

(3) *Test va.dity.*—Cross-validation results, based on pilots tested in advanced and transition training with the satisfaction score as the criterion, are given in table 26.39.

TABLE 26.39.—*Product-moment correlations of Inventory of Experiences, Attitudes, and Interests CE612AX2, with satisfaction scores as the criteria, based on pilots tested in advanced and transition training*

Group	Sample	Key	N	*r*
Bomber pilots	Odds	Evens	901	0.38
Do	Evens	Odds	916	.47
Fighter pilots	Odds	Evens	797	[1]−.34
Do	Evens	Odds	843	[1]−.30

[1] Since the test was keyed to yield positive coefficients against the criterion of bomber-satisfaction, these negative correlations are in the expected direction.

Evaluation.—This test appears to have a quite satisfactory validity. The cross-validation is more a test of internal consistency, however, than it is a genuine test of validity. Since the requirements used in making up the scoring keys were stringent, and since both keys were valid, the indications are that the instrument might be of distinct value in assigning pilots to specialized training. Further evidence supporting this conclusion is the fact that some of the items in part II that have high positive phis are items appearing in the Aviation Preference Check List which significantly discriminate between above-average bomber pilots in transition and above-average fighter pilots in single-engine advanced training.

It must be realized that the data obtained for this inventory were based upon pilots who were in ansition and advanced schools. This is a highly selected group, in that it does not include those pilots who were eliminated at the primary or basic level. In addition, pilots in transition and advanced schools, for the most part, have already decided on the basis of actual experience with one or the other type of training which type of training they prefer, further biasing the group. It is possible that similarly good results might not be obtained if the inventory were administered at a much earlier stage, i. e., during the classification period, or at primary school and then validated on a later-obtained criterion.

Specialization Preference Inventory, CE610A [18]

This inventory was constructed on the assumption that preferences for certain activities and likes and dislikes would differentiate potentially successful fighter and bomber pilots.

Description.—An attempt was made in this inventory to pair systematically some of the alternatives of every category with an equal number of alternatives from other categories. These categories are: intellectual *v.* nonintellectual, mechanical *v.* nonmechanical, routine *v.* nonroutine, social *v.* nonsocial, responsibility *v.* nonresponsibility, social values *v.* economic values, etc. This inventory differs from previous ones in that military activities were matched extensively with civilian activities.

(1) *Internal characterstics.*—The inventory contains 122 items. Two typical items are:

[18] Developed at Psychological Research Unit No. 3. Chief contributors: Lt. John I. Lacey and Lt. Eli A. Lipman.

A. Read a book on gunnery.
B. Read a book on propaganda methods.

A. Building wooden cabinets.
B. Interview job applicants.

(2) *Administration.*—The examinee is instructed to answer each item by indicating the one activity of two presented that he prefers. He is further instructed to answer quickly, according to his first impression, avoiding deliberation.

The time limits are as follows: administration time, 2 minutes; testing time, 25 minutes. The test is paced by announcing the number of the question the examinee should have finished at certain times; namely, item 30 when a quarter of the time has elapsed; item 60 when one-half of the time has elapsed, and item 90 when three-quarters of the time have elapsed.

(3) *Scoring.*—An empirical scoring key based on the results of an internal-consistency item analysis was constructed. This key has 71 positively (R) and negatively weighted (W) items. The scoring formula is R−W+40,.

Statistical results.—The data on this test are restricted to a sample of pilots tested in basic flying training, in classes 44H and 44I, by personnel of Psychological Research Unit No. 3.

(1) *Distribution statistics.*—Based on the final empirical key, a sample of 724 classified pilots yielded a mean score of 32.7 and a standard deviation of 16.9.

(2) *Internal consistency.*—An a priori bomber-pilot key was used to score the answer sheets of 750 classified pilots (sample I). This key represented the consensus of 12 aviation psychologists. No item was accepted unless at least 10 of the 12 judges agreed. With the total score as the criterion, an internal-consistency item analysis was completed. The homogeneity of the items is indicated by a mean phi of 0.27, a standard deviation of 0.13, and a range of values from 0.00 to 0.53. These statistics are based upon analysis of the highest 27 percent and the lowest 27 percent in total score.

A new key was derived from this analysis. The answer sheets of a new sample (II) of 750 cases and of the previous sample (I) were then scored, using the new key. The data in table 26.40 are based on analysis of the responses of the highest 27 percent and the lowest 27 percent in the new total score.

TABLE 26.40.—*Internal-consistency data for 122 items of Specialization Preference Inventory, CE610.4, using the second key, based on 750 pilots in primary training*

Sample	Mϕ	SDϕ	Range of ϕ	
			Low	High
I	0.31	0.11	0.07	0.58
II	.30	.11	.06	.53

(3) *Reliability coefficient.*—By the odd-even method, an estimated reliability coefficient of 0.78, corrected for length, was obtained. This figure is based on a sample of 724 classified pilots.

(4) *Intercorrelations.*—Since a test of this sort is very likely to measure verbal and other intellectual abilities, a correlation between this test and Reading Comprehension, CI614H, was secured. Based on 507 classified pilots, the correlation is only 0.01, corrected for attenuation in both variables. The test, therefore, obviously has no verbal variance.

Correlations with navigator and pilot stanines are only 0.05 and −0.13, based on classified pilots numbering 505 and 634 respectively.

Evaluation.—This inventory has three advantages over previous inventories. It contains a more systematic pairing and a wider coverage of activities. The items are so presented that the purpose of the inventory is less obvious to the examinee.

A review of the items that are highly consistent with the total score in both analyses reveals patterns of interest similar to those empirically evolved from the Aviation Preference Check List and the Satisfaction Test. In table 26.41, a sample is given of these items, together with analogous items that were valid in the Aviation Preference Check List and Satisfaction Test, CE409B. This points strongly to the conclusion that Specialization Preference Inventory could serve as a useful instrument in assorting pilots for specialized training.

Its relatively high internal consistency, unusual for a test of this nature, its low correlations with pilot and navigator stanines, and its lack of correlation with a reading-comprehension test, all indicate the potential usefulness of the inventory.

TABLE 26.41.—*Comparison of some items in Specialization Preference Inventory, CE610A, Aviation Preference Check List, and Satisfaction Test, CE409B*[1]

Specialization preference Inventory items	Aviation preference check list or satisfaction test items
7. A. Perform some types of technical work on a construction project. *B. Have complete administrative responsibility for a construction job.	53. A. Repair a motor. *B. Tell others how to.
19. A. Perform acrobatics in a basic training plane. *B. Complete an instrument flight in an advanced training bomber.	10. A. Acrobatics. *B. Instrument flight.
54. *A. Direct and supervise ground-school curriculum. B. Be in charge of cadet athletic activities.	101. *A. Ground school instructor. B. Physical training instructor.
70. *A. Use maps and landmarks to plot position. B. Be the tail gunner on a B-17.	As an enlisted member of a combat crew, I would get more satisfaction from being a 18. *A. Radio operator. B. Tail gunner. 25. *A. Flight engineer. B. Tail gunner.
76. *A. Be an honor student in a college class. B. Win a varsity letter.	30. *A. Be a famous artist? B. Be a famous football star?
114. *A. Adjust mental difficulties of people. B. Operate a steam locomotive.	81. *A. Have control of other people. B. Have control of a machine (like a plane).

[1] Asterisks indicate items positively weighted for bomber pilots.

Specialization Interest Inventory, CE609A [14]

The rationale for this test is the same as for Specialization Preference Inventory, CE610A.

Description.—This inventory is patterned after the Strong Vocational Interest Blank. As such, it covers an even wider range of interests, both occupational and scholastic, than the Specialization Preference Inventory, CE610A.

(1) *Internal characteristics.*—There are 250 items divided into three parts. Part I contains 185 occupations listed alphabetically, such as auditor, advertiser, architect, astronomer, etc. The examinee must respond to each item in one of three ways—like, indifferent, or dislike. Part II contains 35 items listing school subjects, such as, foreign language, social science, philosophy, etc. Again the examinee must respond to each with like, indifferent, or dislike. Part III has 30 items listing sports and positions on teams (catcher, pitcher, etc.). In each of these items the examinee selects the one he enjoys, or thinks he would enjoy, most and the one he enjoys, or thinks he would enjoy, least. An example of this type of item follows:

Of the following positions in football, which one would you enjoy most?
Of the following positions in football, which one would you enjoy least?

A. End.
B. Guard.
C. Center.
D. Quarter or half-back.
E. Fullback.

(2) *Administration.*—The following sentences are excerpts from the directions:

> Indicate after each occupation listed below whether you would like that kind of work. You are not asked if you would take up the occupation permanently, but whether you would enjoy the kind of work, regardless of any necessary skills, abilities, or training which you may or may not possess * * *
>
> Work rapidly. Use only your first impressions in answering. Answer all the items.

The time limits are: testing time, 28 minutes; administration time, 2 minutes. The examinees are paced by announcing the number of the question the examinee should have finished, i. e., items 63, 126, and 188 at the end of 7, 14 and 21 minutes, respectively.

(3) *Scoring.*—No key is available.

Statistical Results.—No statistical data are available.

Social Concepts, CE512A [15]

This test was constructed for the purpose of determining whether or not successful and unsuccessful air-crew trainees differ in their social beliefs and attitudes.

[14] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. John I. Lacey and Lt. Eli A. Lipman.
[15] Developed at Psychological Research Unit No. 1.

A number of studies of fear in combat, e.g., that of Dollard, indicates that proper motivation of the soldier is an important factor in overcoming fear. Proper motivation also enables the soldier to adjust to new situations and to cope with the deprivations involved in his experience. As belief in the cause for which one is fighting was known to be an important element in military motivation, a test of social concepts and of social morale was constructed.

Description.—This test may be viewed either as a measure of the individual's understanding of or his attitudes towards social problems. Since both views can be assumed to have some bearing on motivation, it was not considered important to distinguish between measures of understanding and measures of attitudes. The method of responding to the statements (true or false, acceptable or unacceptable) does make it possible, however, for the examinee to indicate whether he believes he is dealing with facts or with opinions.

(1) *Internal characteristics.*—The "Test of Social Understanding" developed by the Cooperative Study in General Education (2) was examined and its contents modified in order to make the items suitable for aviation students. One-hundred forty-six items were constructed. A few typical items are:

A. Everybody has an equal chance in America.
B. If we could pass the right laws, we could solve our social problems once and for all.
C. The United States has nothing like social classes.
D. The people who complain about an unfair press are free to start a paper of their own.
E. There never was a modern war that wasn't started by the bankers and munitions-makers.
F. Discussing social issues does not help to solve them.

(2) *Administration.*—The following are excerpts from the directions to the Social Concepts Test:

The statements in this test are those which are frequently heard in the everyday remarks people make * * *

If the statement is true, blacken space A * * *

If the statement is false, blacken the space under the letter B, etc.

If the statement is neither true nor false, but you agree with its point of view, i. e., have a preference for it, blacken the space under C.

If the statement is neither true nor false, but you disagree with its point of view, i. e., do not have a preference for it, blacken the space under the letter D.

If none of these ways describes your reaction to the statement, you may blacken the space under the letter E.

No time limits are set.

(3) *Scoring.*—This scoring formula is R−W+20, in which R refers to the positively weighted responses and W refers to the negatively weighted responses.

Statistical results.—(1) *Item validity.* After dividing a sample of 1,014 answer sheets into two random halves, the responses to the items were correlated with the primary pilot graduation-elimination criterion. The distributions of the phi coefficients are shown in table 26.42. Items are not included if one response is chosen by more than 90 percent or less than 10 percent of the examinees.

TABLE 26.42.—*Distribution of validity phi coefficients for Social Concepts Test, CE512A, based on pilots in primary training, using graduation-elimination as the criterion*

Phi	Odds[1] *f*	Evens[1] *f*	Phi	Odds[1] *f*	Evens[1] *f*
0.15–0.17	1	1	(−0.03)–(−0.01)	104	95
0.12–0.14	7	3	(−0.06)–(−0.04)	72	88
0.09–0.11	20	16	(−0.09)–(−0.07)	31	38
0.06–0.08	37	40	(−0.12)–(−0.10)	9	11
0.03–0.05	97	88	(−0.15)–(−0.13)	3	0
0.00–0.02	128	114	Total	509	494

[1] $N_t = 507$, $p_g = 0.78$. In class 44K; tested at Psychological Research Unit No. 1 in December 1943.

In interpreting these phi coefficients, it can be said that for an N of 507, a phi of 0.09 is significant at approximately the 5 percent level, and a phi of 0.11 is significant at the 1 percent level of confidence. In the odds sample, 32 phis exceed the 5 percent level of confidence and 20 exceed the 1 percent level. In the evens sample, the corresponding figures are 39 and 11. There was almost no correlation between the two sets of significant items, with but three keyed responses in common. For one item a response was keyed positive in one key and negative in the other.

Evaluation.—Since there are only two keyed responses in common in the odds and evens samples, a cross-validation of scores using the two samples probably would show a very low validity for the primary pilot criterion.

Traits measured by this test are believed to be more important in combat than in training. Since combat criteria are difficult to obtain, proper validation of these tests has not been effected.

Survey of Personal Attitudes, CE508B [16]

The purpose of this test is to measure readiness for combat duty in terms of the soldier's affective organization and orientations.

It is assumed in this test that certain established attitudes or affective habits are conducive to efficiency, stability, and endurance in combat situations, while others are not. It is further assumed that by examining the student's likes and dislikes for carefully selected words with appropriate affective connotations, one may obtain a partial measure of his emotional preparedness for combat duty.

Description.—In this test the examinee is presented with series of four words each, designed to have affective connotations. The examinee

[16] Developed at Psychological Research Unit No. 1. Chief contributors: Lt. Vivian E. Fisher, Capt. Donald E. Super.

703320—47—49

indicates which word in each series is most pleasant or most unpleasant. An attempt was made to collect an adequate number of the following classes of words:

(1) Those which have had their unpleasant implications enhanced or colored by the war situation.

(2) Those which have had their pleasant implications enhanced or colored by the war situation.

(3) Those which have not had their implications affected by the war situation.

The words are grouped in units of four in such a way that: (1) each unit contains only words of apparent equal pleasantness or unpleasantness, and (2) each unit contains both words which have had their affective coloring changed or enhanced by the war situation and words which have not been affected.

(1) Internal characteristics.—The test has two parts. Part I consists of 60 groups of words which have unpleasant associations or meanings for most persons. One such group of words, for example is: "disease, slink, selfishness, ugly." Another group: "Jerry, German, Hun, Nazi." Part II of the test consists of 60 groups of words which have pleasant associations or meanings. Examples of these items are: "freedom, power, peace, ability;" and "airspeed, velocity, speed, rate."

(2) *Administration.*—The following are excerpts from the directions to part I:

> The purpose of this test is to learn what words are most unpleasant to you and what words are most pleasant. Experience has shown that one's feelings toward various aspects of life have a bearing on success in cadet training. There are no right or wrong answers * * * You are to read the words carefully and indicate under the appropriate letter on the answer sheet the one word in each group which you find is most unpleasant. If no word in a group is unpleasant, then indicate the word which comes nearest to being unpleasant. Be sure to indicate one word and only one in each group of four words.

In part II, the examinee indicates the one word which he finds to be most pleasant.

(3) *Scoring.*—An a priori key was constructed in the following manner. Positive weights were given to those responses which were thought to evidence readiness for combat. For example, in part I in the series "to injure, to hurt, to wound, and to harm," *to wound* was weighted positively. In part II, in the series "letters, postcards, correspondence, and communications," *letters* was weighted positively. The total score is the sum of the positively weighted responses.

Statistical results. (1) *Distribution statistics.*—Distribution statistics are given in table 26.43.

(2) *Reliability coefficient.*—An estimate of reliability estimated by correlating part I with part II is 0.14. This low correlation indicates either that the two parts do not measure the same thing or that the part scores themselves are unreliable.

Table 26.43.—*Distribution constants for Survey of Personal Attitudes, CE508B, based upon 870 classified pilots*[1]

Part	M	SD
I	16.1	4.4
II	19.2	5.9

[1] Tested in June 1944 at Psychological Research Unit No. 1.

(3) *Test validity.*—Validation results for both parts of the test are given in table 26.44.

Table 26.44.—*Validity data for Survey of Personal Attitudes, CE508B based upon pilots in primary training (N_t=870; p_s=0.79)*[1]

Part	M_s	M_e	SD_e	r_{bis}	$_cr_{bis}$[2]
I	16.08	16.13	4.40	−0.01	−0.04
II	18.97	19.36	5.86	−.04	−.05

[1] Tested in June 1944 at Psychological Research Unit No. 1.
[2] Assuming an unrestricted stanine standard deviation of 2.00.

Evaluation.—This technique designed for measuring "readiness for combat" did not discriminate between pilot graduates and eliminees, at least in primary training. This may be due to a number of reasons. While good combat pilots are aggressive, self-assertive, and happy-go-lucky, these traits may not be susceptible to measurement by a comparatively subtle testing instrument. A second defect might be in using the a priori method of keying words that have varying connotations. If an empirical key were established, this test might prove useful in testing the original hypothesis. Combat criteria, however, are needed for item and score validation of a test of this type.

Inventory of Attitudes, CE518A[17]

This test is designed to measure certain personality traits that are believed to be conducive to the development of psychoneuroses, particularly under near-combat and combat conditions.

Description.—A general orientation to the selection of items was effected by using a digest of personality traits gleaned from "War Neuroses in North Africa" (3) by Grinker and Spiegel and from other field studies.

(1) *Internal characteristics.*—The test is divided into three parts, Part I contains 91 items requesting the examinee to express his opinions of the importance or significance of different combat or near-combat duties and responsibilities. Two examples from part I are:

With regard to combat assignments, my concern as to whether I will have competent leadership from superior officers:

A. Is very great.
B. Is considerable.

[17] Developed at Psychological Research Unit No. 1. Chief contributors: Lt. Vivian E. Fisher, Capt. Donald E. Super.

C. Is slight.
D. Is absent.

I believe the war is hardest on:
A. The fighting men.
B. Parents of fighting men.
C. Children of fighting men.
D. Wives and sweethearts of fighting men.

Part II includes 85 items requesting the examinee to give a direct evaluation of his personality. Typical items from part II are.

If one of my superiors keeps picking on me, I shall:
A. Tell him exactly what I think.
B. Tell my squadron mates what I think of him.
C. Say nothing, but blow off steam some other way.
D. Try not to be disturbed by the matter.

When I fall in love or make friends, the relationships can best be described as:
A. Very intense and short lived.
B. Very intense and long lived.
C. Casual and short lived.
D. Casual and long lived.

The 48 items in part III require the examinee to give his judgment of his friends' or family's evaluation of his personality. Typical items of part III are:

My friends and/or family seem to think I am inclined:
A. Never to give up.
B. To give up only after great effort.
C. To give up easily.
D. To give up very easily.

My instructors usually seem (seemed) to like me:
A. Very well.
B. Fairly well.
C. Very little.
D. Not at all.

(2) *Administration.*—The following are excerpts from the directions:

In this booklet you are asked for certain information about some of your attitudes and opinions. This is not a test in the ordinary sense; there are no right answers except the ones which most truly reflect your own particular attitudes and opinions.

To a large extent your success in flying depends on how well you are understood by those in charge. All of the information asked for in this booklet is for the purpose of aiding your superior officers in understanding you. It is to your own advantage, therefore, to indicate your answers to the items in this booklet as carefully, completely, and honestly as you can.

(3) *Scoring.*—There is no a priori scoring key. An attempt, described below, was made to develop an empirical key.

Statistical results.—Only item validity data are available for this test, based upon pilots originally tested in May 1944 at Psychological Research Unit No. 1.

(1) *Item validity.*—After dividing a sample of 752 answer sheets into two random halves, the responses were correlated with the primary pilot

graduation-elimination criterion. Results are shown in table 26.45. Items are not included if one response is chosen by more than 90 percent or less than 10 percent of the examinees.

TABLE 26.45.—*Distribution of validity phi coefficients for Inventory of Attitudes, CE518A, based on pilots in primary training, using graduation-elimination as the criterion*

Phi	Odds[1]	Evens[1]	Phi	Odds[1]	Evens[1]
	f	f		f	f
0.16–0.19	3	6	(—0.08)-(—0.05)	88	89
0.12–0.15	17	13	(—0.12)-(—0.09)	34	42
0.08–0.11	60	47	(—0.16)-(—0.13)	8	10
0.04–0.07	93	101	(—0.20)-(—0.17)	3	2
0.00–0.03	129	131	Total	572	586
(—0.04)-(—0.01)	137	145			

[1] $N_e=376$, $p_e=0.80$.

In interpreting these phi coefficients, it can be said that for an N of 376, a phi of 0.10 is significant at approximately the 5 percent level, and a phi of 0.13 is significant at the 1 percent level of confidence. In the odds sample, 77 phis exceed the 5 percent level of significance and 23 exceed the 1 percent level. In the evens sample, the corresponding figures are 83 and 20.

Evaluation.—An examination of the items that were keyed in the same direction on both samples does not reveal any pattern of behavior that would differentiate successful from unsuccessful pilots at the primary training level. Further validation is necessary using combat success or failure as the criterion, since this test was constructed on that premise.

Conduct of the War Test, CE520A [18]

This test is designed to measure the extent of the examinee's information concerning the conduct of the war; that is, military events, methods, objectives, and principles. This procedure is based on the hypothesis that men who are motivated by a patriotic interest in the war and by pride in military achievement will acquire more information concerning the conduct of the war. It is believed important to measure military motivation, since aviation students go through a long and strenuous training period, followed by the dangers of combat.

Description.—One hundred twenty-five items were constructed for this test based on political and military events that took place during the war period, 1939 through 1944. Some aviation-information items are included also.

(1) *Internal characteristics.*—The test consists of information items such as the following:

Germany invaded Poland in:
- A. October, 1938.
- B. January, 1938.

[18] Developed at Psychological Research Unit No. 1. Chief contributors: Sgt. Gerald S. Blum, Sgt. Michael Green, and Capt. Donald E. Super.

C. September, 1939.
D. December, 1939.
E. Don't know.

Secretary of Navy Knox was succeeded by:
A. Patterson.
B. Welles.
C. Forrestal.
D. Wickard.
E. Don't know.

(2) *Scoring.*—The scoring formula for this test is R−W/3. For validation purposes, rights and wrongs were scored separately.

Statistical results.—The data that follow are for examinees tested at Psychological Research Unit No. 1 in the period June to August 1944.

(1) *Distribution statistics.*—An example of distribution statistics obtained on this test are given in table 26.46.

TABLE 26.46.—*Distribution constants for Conduct of the War, CE520A for a sample of 673 classified pilots in primary training*

Score	M	SD
Rights	41.0	13.0
Wrongs	75.4	15.4

(2) *Item validity.*—After dividing a sample of 656 answer sheets into two random halves, the responses to the items were correlated with the primary pilot graduation-elimination criterion. Results are shown in Table 26.47. Items are not included if one response is chosen by more than 90 percent or less than 10 percent of the examinees.

TABLE 26.47.—*Distribution of validity phi coefficients[1] for Conduct of the War Test, CE520A, based on pilots in primary training, using graduation-elimination as the criterion*

Phi	Odds[1] *f*	Evens[1] *f*	Phi	Odds[1] *f*	Evens[1] *f*
0.20–0.23	0	2	(−0.04)–(−0.01)	87	90
0.16–0.19	3	2	(−0.08)–(−0.05)	64	57
0.12–0.15	11	9	(−0.12)–(−0.09)	28	30
0.08–0.11	41	37	(−0.16)–(−0.13)	7	10
0.04–0.07	89	80	(−0.20)–(−0.17)	4	2
0.00–0.03	86	105	Total	420	424

[1] $N_e = 328$, $p_e = 0.68$.

In interpreting these phi coefficients, it can be said that for an N of 328 a phi of 0.11 is significant at approximately the 5 percent level, and a phi of 0.15 is significant at the 1 percent level of confidence. In the odds sample, 37 phi's exceed the 5 percent level of significance and 10 exceed the 1 percent level. In the evens sample, the corresponding figures are 43 and 9.

(3) *Test validity.*—Validation data were obtained for a sample of 673 classified pilots in primary training. The results are shown in table 26.48.

TABLE 26.48.—*Validity of Conduct of the War, CE520A, based on graduation-elimination from primary pilot training* (N_t=673, p_g=0.68)

Score	M_g	M_e	SD_t	r_{bis}	${}_cr_{bis}$[1]
Rights	40.33	41.54	13.05	−0.06	−0.03
Wrongs	74.90	75.82	15.38	−.04	−.04

[1] Assuming an unrestricted stanine standard deviation of 2.00.

Evaluation.—The hypothesis that men who are motivated by a patriotic interest and have therefore obtained more knowledge concerning the conduct of the war will be more successful in air-crew training, is rejected. This conclusion is based on the negative correlations between the scores of 763 pilots in primary training and the graduation-elimination criterion. This is in line with the validities of two achievement tests: Geography, AS104 (X2), and History, AS153 (X3), both of which were slightly negative. Statistical analysis of the questions in the test reveals that they were too difficult, even for persons familiar with military and political events.

Home Front Attitude Inventory, CE446A[19]

The purpose of this test is to measure a certain aspect of predisposition to combat neurosis. Lack of confidence in the home front, according to Grinker and Spiegel (3), was frequently observed in victims of mental disorders in the Tunisian campaign. Reports from combat theaters repeatedly cite evidence of the adverse effect on military morale of incidents at home that lead soldiers to believe that civilians are not doing their part in the war. This test is an attempt to explore this emotionally significant area which is not adequately covered by any tests in the battery.

Description.—The test consists of 100 items, each of which is a two- or three-sentence description of the behavior of a fictitious civilian in a specific situation. The examinee is asked to indicate for each sample of civilian behavior whether it is "very common," "fairly common," "fairly uncommon," or "very uncommon." He is also asked to indicate whether he believes that the civilian conduct is "very justifiable," "fairly justifiable," "fairly unjustifiable," or "very unjustifiable." Responses to the test items should indicate individual differences regarding (1) what is felt to be the typical wartime behavior of civilians; (2) how various wartime behavior tendencies of civilians are evaluated by aviation students; and (3) the degree of confidence which is held regarding the support of the war effort by civilians.

(1) *Internal characteristics.*—Test items cover such areas of activity as the economic life of the country, individual economic interests, sexual needs and romantic interest, political actions, degree of civilian sacrifices,

[19] Developed at Psychological Research Unit No. 1.

etc. There are three parts to the test. Part I has 37 items, part II, 37 items, and part III, 26 items. Three typical items are:

Mr. Z was the leader of a union in which members were receiving more pay than ever before in their lives. There were some causes for grievance among the workers, but he smoothed them over so as to avoid interfering with the war effort.

Mrs. B had always been a little afraid of the sight of blood and was somewhat timid about giving to the Blood Bank. However, when she was reminded that her blood might save a soldier's life, she agreed to donate.

G was in love with a beautiful girl when he was drafted. The letters which he received from her continued to be warm and affectionate even though she had fallen in love with someone else. She felt it would be unpatriotic to "let him down while he was out there fighting."

(2) *Administration.*—The following are excerpts from the directions:

This is a test of your judgment about events which are taking place among civilians and/or which may take place in the future.

Each paragraph is assigned four different numbers on your answer sheet. The corresponding numbers will be in your test booklets to the left of each paragraph. Which number you use on the answer sheet will depend upon your judgment of how *common* the described behavior (thinking, feeling, talking, or acting) is. The first of the numbers will be used if you think it is very common; the second number will be used if you think it is fairly common; the third numbers will be used if you think it is fairly uncommon; and the fourth number will be used if you think it is very uncommon.

After you have decided which number to use, you will have to decide which letter to mark after the number you have chosen. The letter you mark will depend on your judgment of how justifiable such civilian behavior is. You will mark the A space if you think it is very justifiable, the B space if you think it is fairly justifiable, and the C space if you think it is fairly unjustifiable, and the D space if you think it is very unjustifiable. The E space will never be marked.

(3) *Scoring.*—The original system was such that each of the 16 alternatives to an item of the test was given weight in the scoring. It assumes that all items are equally discriminating and that a high positive correlation between judgments of frequency of occurrence and judgments of justifiability is indicative of optimum morale. Weights, as follows, were assigned to responses to each item.

Item	*A*	*B*	*C*	*D*
1	8	6	4	0
2	6	5	3	4
3	4	3	5	6
4	0	4	6	8

Because of the great practical difficulties with this scoring method, it was dropped.

Statistical results.—The only statistical data reported are item-validity studies, based upon the responses of examinees tested at Psychological Research Unit No. 1 in August 1944.

(1) *Item validity.*—After dividing a sample of 740 answer sheets into two random halves, the responses to the items were correlated with the primary pilot graduation-elimination criterion. With this size of sample,

it is feasible to obtain item validities for judgments of commonness of behavior only, since only a small proportion of examinees selects any one of the 16 responses concerning justifiability of behavior. The distributions of the phi coefficients are shown in table 26.49. Items are not included if one response is chosen by more than 90 percent or less than 10 percent of the examinees.

TABLE 26.49.—*Distribution of validity phi coefficients for Home Front Attitudes, CE446A, based on pilots in primary training, using graduation-elimination as the criterion*

Phi	Odds[1]	Evens[1]	Phi	Odds[1]	Evens[1]
	f	*f*		*f*	*f*
0.24–0.27	0	2	(—0.04)-(—0.01)	47	55
0.20–0.23	0	0	(—0.08)-(—0.05)	42	32
0.16–0.19	1	3	(—0.12)-(—0.09)	11	11
0.12–0.15	3	6	(—0.16)-(—0.13)	1	3
0.08–0.11	17	20	(—0.20)-(—0.17)	0	1
0.04–0.07	33	34	(—0.24)-(—0.21)	1	0
0.00–0.03	57	43	Total	213	210

[1] $N_g=370$, $p_g=0.72$.

In interpreting these phi coefficients, it can be said that for an N of 370 a phi of 0.10 is significant at approximately the 5 percent level, and a phi of 0.13 is significant at the 1 percent level of confidence. In the odds sample, 16 phis exceed the 5 percent level of confidence and 4 exceed the 1 percent level. In the evens sample, the corresponding figures are 30 and 14. For no item did the two keys agree.

Evaluation.—Since little empirical evidence is available, there are few conclusions that can be stated about this test. What little evidence does exist, does not support the original hypothesis. The novel feature of this instrument—the double response and scoring based on agreement of two responses—could be suggestive of other uses to which these devices might be adapted. This scoring system, however, is impractical, if not prohibitive, for use with the machine-scoring of large numbers of answer sheets.

EVALUATION OF MOTIVATION MEASURES

Research in the field of motivation has verified previous results and has revealed much that was formerly unknown. The two methods of approach utilized in air-crew classification, one subjective and the other objective, provide an interesting contrast, both of techniques and of results.

Analysis of the voluminous data obtained for the Preference Blank shows conclusively that expressed preferences and strength of interest for pilot or bombardier training do not correlate appreciably with either the pilot or bombardier criteria or with the classification tests that have high correlations with the criteria. On the other hand, preferences and strength of interest for navigator training do correlate significantly and substantially with the criterion, and, in addition, with the navigator stanine, and with individual tests that correlate well with the navigator criterion. This may mean that those interested in navigation also have insight into their abilities and temperament and appreciate the demands of the job they desire to perform.

More objective instruments, such as the Satisfaction test and the two aviation-interest tests that were introduced into the classification battery—General Information and Biographical Data—have higher correlations with the pilot criterion. This substantiates the conclusion that objective techniques are needed for an effective assessment of an examinee's interests.

While not enough data are available to make a definite generalization on the point, the available results from the attitude and interests inventories suggest that the factors that primarily account for fighter-pilot success are also factors that were found to account for success in primary and basic training. These factors are: experiences in mechanics and active sports, general tendencies toward recklessness, adventurousness, extravagance, and a devil-may-care attitude. It is probable that the bomber pilot requires some of these traits; but that, in addition, he must possess more special social intelligence (leadership), and be more conscientious, methodical, thorough, cautious, and persevering. There are also data indicating that pilots assigned to heavy-bomber training have more interest in navigator training than do fighter pilots.

Traits measured by the combat-readiness tests are expected to be more important in combat than in training. Since combat criteria are difficult to obtain, proper validation of these tests has not been effected. Therefore, it has not been possible to ascertain their usefulness.

BIBLIOGRAPHY

(1) Dollard, John. *Fear in Battle,* New Haven: The Institute of Human Relations, Yale University, 1943.

(2) Cooperative Study in Education, Chicago: American Council on Education. Vols. 2–4, Nos. 12, 1, respectively.

(3) Grinker, R. R., and Spiegel, J. P. *War Neuroses in North Africa,* prepared and distributed for The Air Surgeon, Army Air Forces, by the Josiah Macy Jr. Foundation, New York, September 1943.

CHAPTER TWENTY-SEVEN

Biographical Data[1]

INTRODUCTION

Research by Civil Aeronautics Administration

The decision of Army Air Forces psychologists to construct a biographical information blank for air-crew candidates was influenced by studies in this area sponsored by the Civil Aeronautics Administration (2, 3). Prior to Pearl Harbor, research was conducted at Tulane University and the University of North Carolina—with data gathered at the Naval Air Station, Pensacola, Fla.—under the direction of Dr. H. M. Johnson (2). Statements of biographical information and of interests obtained through personal interviews with 480 students at the Naval Air Station and their responses to the 125 items of the Bernreuter Personality test were analyzed. The answers to the Bernreuter questionnaire were regarded as biographical information in the sense that they purported to indicate some of the individual's likes and dislikes, habits of social adjustment, interests, etc.

An item validation of the Bernreuter questionnaire revealed nine potentially useful items. When weighted in the best possible manner, these items yielded a multiple correlation of 0.35 (unshrunken) with the pass-fail criterion for pilot training. The nine items were combined with other items of biographical information gathered in the personal interview. The latter were more objective, concerning facts about education, religion, occupations, and athletics. The coefficient of multiple correlation between the 15 most predictive biographical-information items (including the seven most predictive Bernreuter items), optimally weighted, and the pass-fail pilot criterion for a sample of approximately 280 pilots was 0.51. In obtaining this figure, shrinkage was estimated, but the items had not been administered to a new sample. Experience in the Army Air Forces has shown the utmost importance of the validation of any weighted composite score on an entirely new sample of adequate size.

Studies in Industry

In addition to the promising results of these early CAA studies, there had been experimentation in the industrial field which tended to confirm belief in the value of biographical information for predicting success in a particular area of activity. For example, studies involving life-insurance

[1] Written by Tech./Sgt. Sanford J. Mock.

company employees revealed that certain items of information from personal and occupational histories were predictive of success or failure in specific life-insurance jobs. This conclusion was reached through an empirical evaluation of application forms and of interview data.

An Extreme View of the Biographical Approach

An extreme point of view expressed by Guthrie (1), can be introduced here to support the biographical-data type of approach, without necessarily subscribing to his theory of personality or his convictions concerning the usefulness of specific versus general information as a basis for prediction. He asserts that:

> The useful categories in describing what may be expected of a man include his skills which can be often readily measured. They include his types of adjustment described in terms of the situations to which he has been exposed. He is a hardened campaigner, an experienced broker, an experienced carpenter; he has been for 10 years a head waiter. If we know these types of experience, we may safely assume that he has learned skills that meet the problems of these trades and occupations.

In speaking of application blanks, the same author points out that

> * * * we may require the statements of sponsors as to his (applicant's) introversion or extraversion, his general honesty, his loyalty, his industry. But the useful information on this blank is more likely to be his past record of occupation, his specific skills, his financial status, his marital and police record. His past affiliations, political and religious, offer better and more specific predictions of his future than any of the traits that we usually think of as personality traits (1, p. 66).

Job Analysis Data

An analysis of the jobs performed by pilot, bombardier and navigator encourages the hypothesis that persons with certain types and combinations of educational, physical, social, recreational, and occupational histories would be likely or unlikely to possess the necessary characteristics for air-crew success.

The task of the pilot.—As an example, consider one aspect of the pilot's job—manipulation of controls. The flight controls include the stick, rudder control, flap control, throttle, trimming controls, propeller pitch, and brakes. Operation of these controls requires coordinated movements of arms or legs or both. Changing sensory stimuli must be followed by proper motor responses coordinated with them. We know that although both large and small muscle groups are used, the large muscle groups are here most important. We also know that the movements are varied, depending on the immediate situation; that they may be planned and deliberate but at times must be prompt and automatic.

It is quite reasonable to suppose that individuals who have successfully engaged in activities involving similar motor patterns will have superior aptitude for piloting. Even with an incomplete analysis of the pilot's motor functions, we are given sufficient leads for framing questions, asking accordingly for athletic and occupational experience. For example, the man who is proficient in tennis, basketball, motorcycling, or ice hockey might well be expected to have certain qualities of motor co-ordination and control which may assist him substantially in learning to be a pilot. The same can be said about the individual who, in civilian life, was an acrobat or operator of a complex machine. such as a crane.

The tasks of bombardier and navigator.—The bombardier and navigator, in contrast to the pilot, use small-muscle groups predominantly. The bombardier in operating the bombsight and releasing bombs must make precise movements involving fine adjustments. Precise eye-hand coordination is required. In using many of the navigational instruments, eye-hand or eye-finger coordination is required. For example, in a cramped position, and holding the sextant with the left hand, the navigator must center a comparatively sensitive bubble in its chamber and, using his right-hand fingers on the adjustment knob bring the star under observation into the bubble. Furthermore, he must clock his recorder at the instant the bubble is centered and the star is in the center of the bubble. This description augurs a history of activities demanding motor coordination different from that of the potentially successful pilot. Again, it would seem reasonable that an individual proficient in motor patterns similar to those of navigator and bombardier would have a superior aptitude for these positions.

General experience and education.—We can generalize this conception. Possibly, it would be fruitful to delve into every phase of a candidate's biography that seems to have a connection, direct or indirect, with the characteristic habits required for success in the various air-crew jobs. On this basis, for example, questions were inserted in the biographical data blank which sought to reveal the examinee's mechanical interest and experience, because it was known that the pilot must have an understanding of mechanical principles and must be able to work with mechanical devices. Questions were asked about mathematical knowledge, extent of education, and reading habits, because it was known that the navigator had to have a high degree of numerical proficiency; and it was assumed that he was, on the whole, more intellectually inclined than the pilot. These examples illustrate the approach that was based on the hypothesis that specific aptitudes can be inferred from a knowledge of experience and background.

Summary

The above discussion implies the premise that the habits, motor or intellectual, that an individual has learned in the past will, by transfer, aid him in learning air-crew duties. This is no doubt true. But it does not preclude another premise, that what he has done before successfully indicates constitutional aptitudes for learning those habits, and interest in activities, vocational or avocational, which yielded him satisfactions because he was ready to do well in them.

To summarize, it can be said that aviation psychologists began construction of a biographical information blank because of: (1) Previous studies in this area sponsored by the Civil Aeronautics Administration, (2) experience in industry, (3) job analysis, and (4) armchair reasoning.

Biographical Data Blank, CE602A [2]

This is the first form of the Biographical Data Blank.

Description.—The items relate to individual interests, attitudes, and background. The first group of 20 items asks for information about the examinee's home and personal history. These questions are designed to reveal such facts as father's education and occupation, parents' national stock and marital and financial status, the population size and general location of the area where the examinee lived, the examinee's religion, the extent and emphasis of his education. The next 10 questions ask the examinee to rate, according to a graduated scale, his interest in various subjects studied in school. Questions 31 to 61 ask him to indicate the extent of his interest in certain activities such as art, literature, music, radio, science, mathematics, hiking, smoking, etc. The next 12 items ask for the degree of proficiency he possesses in sports. Then nine items elicit information about his previous employment and occupation and his occupational preference. Questions 86 to 95 concern his military experience, including previous Air Corps jobs. Questions 96 to 116 ask for expressions of preference for military and civilian air-crew jobs. The last group, 117 to 150, consists of positive statements about controversial matters. The examinee is asked to indicate his degree of agreement with the statement. Typical statements are: "Skill is more important than judgment in flying;" "Discipline should be as strict in the Air Corps as in other branches of the army;" "A pilot who has had more than one drink should not fly a plane."

(1) *Internal characteristics.*—Biographical Data, CE602A, contains 150 items divided into the 8 sections described below.

(2) *Administration.*—All items and directions are included in the test booklet. The directions attempt at the outset to change the set of the examinee who is expecting a typical test. An explanation is made of why information about the student's background is important. It is the function of the directions to establish rapport so that honest, straightforward, and cooperative answers will be given. All examinees are allowed to finish the blank, which requires approximately 45 minutes.

Following are the pertinent parts of the directions and an illustrative item from each section:

In this blank you are asked for certain information about your background, your family, your home, your education, your interests, and your attitudes. This is not a test. There are no right answers except the answers that tell the truth about yourself.

[2] Developed at Psychological Research Unit No. 1. Chief contributor: Lt. Col. Laurance F. Shaffer.

To a large extent, your future success as a pilot will depend on how well you are understood by those in charge of your flight training. All of the information asked for has been shown to be related to the proper training of pilots. It is therefore to your own interest to fill out this blank carefully and completely. In no case should any part be omitted * * * Read each question and the answers that follow it. Among the several suggested answers, select the one that best answers the question for you * * * Work as rapidly as you can without being careless. Don't think a long time about questions that ask for your interest and opinions, but record your first impressions rapidly. Everyone should finish the blank, and answer all of the questions * * *

SECTION I

Your father and his parents were chiefly of:

A. American stock.
B. Northern European stock. (English, Irish, Scandinavian, German, etc.)
C. Southern European stock. (French, Italian, Spanish, etc.)
D. Slavic stock. (Russian, Polish, Greek, etc.)
E. Other.

SECTION II

Directions: Indicate how well you like each school subject listed below by blackening the spaces as follows:

A. Liked the subject exceptionally well.
B. Liked the subject somewhat.
C. Indifferent—did not care.
D. Disliked the subject.
E. Never studied the subject.

21. Mathematics.
22. Sciences.
23. History.
24. English; literature.

SECTION III

Rank the following five from A (liked best) to E (liked least):

54. Writing a technical report.
55. Riding horseback.
56. Developing a business system.
57. Repairing a radio.
58. Soliciting contributions for charity.

SECTION IV

Directions: For each activity listed below, you are to blacken a space to indicate how well you perform that activity according to the following scale:

A. Exceptionally well.
B. Well.
C. Fairly well.
D. Poorly.
E. Do not engage in this activity.

Thus if you play football fairly well, blacken the space under C in Row 62; if you do not play football at all, blacken under E.

62. Football, rugby, or soccer.
63. Basketball, hockey, or lacrosse.
64. Baseball or softball.
65. Boxing, wrestling, or water polo.

SECTION V

Directions: Mark any of the following types of work which you have done at any time, and for which you have received remuneration. (More than one may be marked.)

79-A. Manufacturing industries (machine operator, factory hand, textile worker, etc.).
79-B. Technical trades (baker, electrician, radio repairman, etc.).
79-C. Transportation and communication (truck driver, linesman, deckhand, etc.).
79-D. Business trades (store clerk, salesman, agent, window dresser, etc.).
79-E. Public service (fireman, policeman, forest ranger, soldier, etc.).

SECTION VI

Directions: Below are a number of statements about which there are wide differences of opinion. Indicate your personal opinions, whether they agree with those of others or not. Tell how you feel about the statement by blackening one of the spaces as before, according to this scale.

How much flying experience have you had:

91-A. Never been in the air.
91-B. Have flown some in transport ships.
91-C. Fly with friends occasionally.
91-D. Have had some instruction.
91-E. Have soloed.

SECTION VII

Directions: Below are listed five types of military piloting. Choose the one of the five types that you would most like to do, and blacken the space under A for its number. Now decide which you would like next best, and blacken the space under B at the right of its number. In the same manner, indicate your third, fourth and last choices by blackening under C, D, and E. Do not give any two of the five the same rank.

96. Scout observation planes.
97. Light bombers.
98. Pursuit planes.
99. Pilot in command of large bomber.
100. Test pilot.

SECTION VIII

A. Strongly agree.
B. Agree.
C. Undecided.
D. Disagree.
E. Strongly disagree.

117. Almost any normal young man can learn to fly.
118. Most airplane accidents could undoubtedly be prevented if certain inherent structural defects of planes were removed.

Statistical results. Validation data are available for cases tested in January 1942 at Psychological Research Unit No. 1.

(1) *Test validity.*—The answer sheets of 574 graduates and 304 eliminees from primary pilot training (classes 42H, 42I, and 42J) were divided into two equal groups. A key was prepared on the basis of each group, weighting items that differentiated (simple difference in percent-

ages) at the 1 percent level of significance. Cases with previous flying experience were omitted. The key derived from each group was validated against primary graduation elimination in the other group. The validity coefficients (biserials) were 0.43 and 0.36.

(2) *Item validity.*—The validity of responses in this test is indicated by a range of phis from −0.24 to +0.30, with a standard deviation of 0.06. A significant negative phi, of course, is just as important as a positive one of comparable size. The data are based upon the responses of the above-mentioned sample of 574 graduates and 304 eliminees from primary pilot training.

Evaluation.—Validity statistics computed for Biographical Data, CE602A, indicated that the hypothesis that aptitude for flying would be predicted by scoring certain biographical information was justified. The extent of the validity that could be expected consistently was not yet definitely ascertained, but the approach was most promising. The next step was to revise and to construct new items on the basis of the item analysis of CE602A.

Biographical Data Blank, CE602B [8]

This is the second form of the biographical data blank.

Description.—There is a considerable difference in the composition of forms CE 602B and CE602A. Form A contains a mixture of items, including biographical questions of fact, and also questions involving interests and preferences. Form B contains only biographical questions of fact. The emphasis is on "What have you done?" rather than, "What do you like?" There were misgivings concerning the use of subjective judgments as in parts of CE602A, in which intentional or unintentional misrepresentation is possible. It was not claimed that the strictly biographical type of question is completely free from the possibility of falsification, but it was felt that factual statements are less likely to be falsified than statements of opinion. They were also believed to be more answerable, in that an individual knows what events or conditions occurred within his experience, whereas he has not formed definite opinions and attitudes in all areas of life or has not verbalized them sufficiently to report them in a precise manner.

(1) *Internal characteristics.*—This form consists of 125 items. The categories covered include origin and personal history of parents, early home environment of examinee, attitude of parents toward examinee's Air Corps career, subjects studied in school and proficiency attained, hobbies, proficiency in athletics, participation in mechanical and literal activities, occupational experience, and military experience.

(2) *Administration.*—This form is administered in the same manner as CE602A. The total time required is approximately 40 minutes.

(3) *Scoring.*—In April 1942, form CE 602B was administered experi-

[8] Also developed at Psychological Research Unit No. 1. Chief contributor: Lt. Col. Laurance F. Shaffer.

703320—47—50

mentally to a large number of studen' along with the classification battery at Psychological Research Unit No. 3, Santa Ana. Later, 1,882 of these examinees graduated from primary pilot training, while 735 were eliminated (classes 42I, 42J, 42K, and 43A). An item-validation study was made and a pilot key constructed on the basis of correlations of responses with the pass-fail criterion. This key covered 51 responses, 22 of which were weighted plus 1 and 29 minus 1, the algebraic sign being consistent with that of the phi correlation.

The same blank, CE 602B, was administered with the classification battery to 3042 students at Psychological Research Unit No. 1, Maxwell Field, during the period Feb. 12 through Apr. 2, 1942. Of these, 598 who entered pilot training were used in validation studies. The biographical data blank was not referred to as an experimental test, nor treated in any other way that would distinguish it from the classification tests either at Santa Ana or at Maxwell Field. The validation of the blank with these samples, therefore, represents results that might be expected if the blank were administered subsequently for classification purposes.

The key furnished by Psychological Research Unit No. 3 yielded a biserial of 0.34 for the 598 pilot students tested by Psychological Research Unit No. 1. (See table 27.1.) The Santa Ana key was then augmented by including all items (except those dealing directly with actual flying experiences) that showed a difference of 5.6 between the percentage of responses for the graduates and eliminees based on the Psychological Research Unit No. 3 tabulations. For N's of the magnitude involved in this study, a difference of 5.6 was significant at the 1 percent level for a response chosen by half the group. For items chosen by any other proportion of the group, a difference of 5.6 had a higher level of significance. On this basis 21 responses were added to the key, and 3 were deleted, making a total of 69 responses, 37 weighted plus 1 and 32, minus 1. The augmented key yielded a biserial *r* of 0.42 for the same group of 598 pilot students. (See table 27.1.)

Statistical results. (1) *Distribution statistics.*—The distribution of scores in this test is indicated by a mean score of −2.0 and a standard deviation of 5.2. The scoring formula is R−W, in which R and W stand for the positively and negatively weighted responses, respectively. These data are based on the above-mentioned sample of 598 classified pilots scored with the Psychological Research Unit No. 3 key, which contained 22 positive and 29 negative responses. The same 598 classified pilots were scored with the Psychological Research Unit No. 1 augmented key which contained 37 positive and 32 negative responses. The mean score was 3.7 and the standard deviation 7.0.

(2) *Test validity.*—The validity data are summarized in table 27.1. These include results from two other keys developed at Psychological Research Unit No 1 and at Headquarters, Army Air Forces, based on all available data.

TABLE 27.1.—*Validity of Biographical Data, CE602B, for primary pilot training, graduation-elimination criterion*

Key	N_t	r_e	M_e	M_s	SD_t	r_{bis}
PRU No. 3 (51 scored responses) ..	[2]598	0.70	−1.03	−3.94	5.23	0.34
PRU No. 3 (69 scored responses) ..	[2]598	.70	5.33	0.46	4.87	.42
PRU No. 1 (47 scored responses) ..	[3]900	.63	3.76	2.00	4.7	.23
AFTAS[1] (102 scored responses)	[3]900	.63	7.51	4.46	6.6	.29

[1] The Air Surgeon's Office, Headquarters, Army Air Forces.
[2] Same sample, tested at Psychological Research Unit No. 1 from Feb. 12 to Apr. 2, 1942.
[3] Same sample, tested at Psychological Research Unit No. 1 in September 1942. In class 43F.

(3) *Item validity.*—The data are summarized in table 27.2.

TABLE 27.2.—*Item validity data for Biographical Data, CE602B, based on 1,182 graduates and 735 eliminees from primary pilot training*[1]

Group of items	$M\phi$	$SD\phi$	Range of ϕ	
			Low	High
22 responses keyed+1	0.09	0.03	0.03	0.17
29 responses keyed−1	−.11	.03	−.16	−.05

[1] In classes 42I, 42J, 42K, and 43A. Tested at Psychological Research Unit No. 3 in April 1942.

Evaluation.—The predictive value of the biographical-information items was sustained in the second form of the Blank, CE602B. An augmented key based on 69 responses yielded a pilot validity of 0.42. In the preparation of this form, all items of opinion were eliminated, leaving questions of fact only. This apparently did not alter the test validity significantly.

Variations of the test.—Two variations of this form are worthy of mention—CE602B-SA and CE602SAB.

(1) Form CE602B-SA.[4]—This form is a version of CE602B prepared especially for experimental administration in January 1943 to pilots. It contains 115 items, including 20 questions (part II) asking for personal opinions, similar to the final section in the first Blank, CE602A. The scoring formula is R−W+20. It was administered at Psychological Research Unit No. 3 in an experimental battery just before the students left preflight school for primary school.

Statistical results. (1) *Distribution statistics.*—In table 27.3 are presented distribution constants, using three keys.

TABLE 27.3.—*Distribution data for samples of pilots scored with different keys on Biographical Data, CE602B-SA*

Key	Part	N	M	SD
Original key	I	[2]856	20.3	7.8
Do	II	[2]856	20.6	5.1
Stanine key[1]	Total	[3]1,645	19.2	3.2
Validity key[1]	..do	[3]1,617	19.6	4.6

[1] In the stanine key, items were weighted to maximize correlation with stanine and to minimize correlation with the graduation-elimination criterion. The validity key was constructed to accomplish the reverse function. See also pp. ——
[2] In class 43H.
[3] In class 43J.

[4] Developed at Psychological Research Unit No. 3. Chief contributors: Capt. S. W. Cook, Col. J. P. Guilford, Capt. L. G. Humphreys, and Maj. Merrill F. Roff.

(2) *Test validity*.—Validation results are summarized in table 27.4.

TABLE 27.4.—*Validity data for Biographical Data Blank, CE602B-SA for pilot training, graduation-elimination criterion*

Group	Part or key	N_t	p_g	M_g	M_e	SD_t	r_{bis}	$_c r_{bis}$[1]
In primary training	I	[2]856	0.75	21.75	16.20	7.75	0.42	0.45
Do	II	[2]856	.75	20.92	19.84	5.10	.13	.13
Through basic training	I	[2]856	.67	22.05	16.95	7.75	.40	.43
Do	II	[2]856	.67	21.04	19.86	5.10	.14	.15
In primary training	Stanine key	[3]1,645	.62	19.31	19.05	3.17	.05	....
Do	Validity key	[3]1,617	.62	20.62	17.86	4.64	.37	....

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] In class 43J.
[3] In class 43H.

(3) *Item validity*.—Item validity data for 2,500 pilots in primary training are shown in table 27.5. The proportion of this sample that graduated is unknown.

TABLE 27.5.—*Item validity data based on 2,500 pilots in primary training*[3]

Group of responses	$M\phi$	$SD\phi$	Range of ϕ	
			Low	High
Responses keyed + 1	0.04	0.04	−0.06	+0.19
Responses keyed − 1	−.05	.06	−.18	+.13

[3] In class 43H.

Evaluation.—Form CE602B-SA revealed the relative lack of predictiveness of the opinion items, as indicated by biserials of 0.13 and 0.15, as compared with 0.45 and 0.43 for the factual items. Tests with as low validity as for the opinion items often add something to prediction, however, particularly if their contributions are unique. Since this seemed a distinct possibility in the case of these items, further attention was given to them in an enlarged test known as Survey of Aviator Opinion, CE604A (see ch. 25).

(2) Form CE602SAB.[4]—This form contains 22 items, divided into two parts. Part I (Biographical) consists of 12 items, and part II (Opinions and Interests), of 10. All examinees were allowed to finish, which required approximately 7½ minutes. Most of the items were based on ideas presumed to be valid by nonpsychological personnel in the classification section of the Santa Ana Army Air Base Classification Center. These observers believed that certain answers to the included questions would indicate a favorable background for flying success. The questions were concerned primarily with military experience, before and after entering the Air Forces. Following is a typical question:

How much military experience had you had before entering the Air Forces?
- A. Member of an organization that has been overseas under fire.
- B. Member of an organization that has been overseas but not under fire.
- C. Member of an organization that had been alerted for overseas duty.

[4] Developed at Psychological Research Unit No. 3. Chief contributor: Lt. David H. Jenkins.

D. Was in training or on duty in the United States only.
E. Came directly into Air Forces after induction.

Part II asked questions of attitude and interest, but this part differed from the interest sections in the previous blanks in that it was directed almost wholly towards aviation and military interest.

To what extent do you like to work with motors and engines?
A. Prefer it to most other things.
B. Like it as much as most other things.
C. Like it less than most other things.
D. Do not like it at all.

Statistical results.—Biographical Data, CE602SAB, was administered at Psychological Research Unit No. 3 in November 1943 to 1,700 pilots, and the items were validated against graduation-elimination.

(1) *Item validity.*—The distribution of phis is described in table 27.6.

TABLE 27.6.—*Item validity for Biographical Data, CE602SAB, for pilots in primary training,[1] graduation-elimination criterion*

Group	N_t	$\bar{\phi}_t$	SDϕ	Range of ϕ
Pilots in primary training	675	0.89	0.07	−0.13−0.14
Pilots through advanced[2]	638	.77	.07	−.15−.16

[1] In classes 44F and 44G.
[2] Below-average students (who eventually graduated) were counted as eliminees.

Evaluation.—As mentioned previously, the 22 items in Biographical Data Blank, CE602SAB, were based on opinions of classification-section[6] personnel as to what biographical information would be predictive. Statistics are not complete, but the item-validity data indicate a general lack of predictive value of the items as compared with previously validated items.

Biographical Data Blank, CE602D[7]

This is the only form of the Biographical Data Blank that has been in the classification battery, which it entered in July 1943. It is based upon items tried out experimentally in the previous forms. Sixty-five of the items from form CE602B that showed empirical validity for pilot or navigator prediction were incorporated into CE602C. CE602D, the battery form, contains the same 65 items more conveniently arranged for administration.

Description. (1) *Internal characteristics.*—The items of form CE602D are distributed among the following categories: Origin of parents and early home environment, subjects studied in school and proficiency attained, proficiency in athletics, relative frequency of performance of mechanical and academic activities, hobbies, occupational experience, and aviation-training interest.

[6] A nonpsychological agency which made final decisions concerning classification.
[7] Developed by cooperative efforts of Psychological Research Unit No. 1 and Psychological Research Unit No. 3.

(2) *Administration.*— ‘ince it is desired that all examinees finish, the time limit is generously . t at 25 minutes.

(3) *An experiment r administration.*—To determine the probable degree of falsification an' its effect upon pilot validity, form CE602D was administered to equiv . ent groups of students under three sets of instructions designed to ›roduce different degrees of motivation to falsify or not to falsify. One set of instructions was framed with an attempt to minimize the pressur: against falsification, a second set to maximize such pressure; and a third set was identical with the standard instructions for the test.

Following are the salient sentences from the first set of instructions, in which minimal pressure against falsification is exercised:

> Read each question and the answers that follow it. Select the statement that best answers the question for you. Do not try to see all possible interpretations of a question or answer; these are meant to be straightforward questions with simple answers. Use approximate answers when in doubt. Do not quibble over small inaccuracies. Remember that what counts is the general impression made by your answers together. No single response can decide your future.

Following are the pertinent portions of the second set of instructions, in which maximal pressure against falsification is exercised:

> As stated in Army Regulations 380-5, the Army Air Forces attaches great significance to the response you are about to make. Read each question and the answers that follow it. You must be certain to select the statement that best answers the question for you. Do not attempt to answer with the choices which you think will result in your being given the particular assignment you want.
>
> Your signature attests to the truthfulness of your answers. If upon checking your references, we find that you have committed perjury, your entire future in the Air Forces will be endangered. Tests such as this are frequently used to detect perjurers, since it is possible to check on the truthfulness of your answers from other sources. The seriousness of fraud and perjury in the Army is set forth in the fifty-fourth, ninety-third, and ninety-sixth Articles of War. The fifty-fourth Article of War states specifically, "Any person who shall procure himself to be enlisted in the military services of the United States by means of willful misrepresentation or concealment as to his qualifications of enlistment, and shall receive pay or allowance under such enlistment, shall be punished as a Court Martial may direct * * *"
>
> Work as rapidly as you can without being careless. Do not think a long time about questions that call for simple facts, but mark quickly what is true in your case. Everyone should finish the blank and answer all the questions. Omit none. You have 25 minutes. Remember, you are subject to military law. Answer with strict honesty.

Following are the sections from the standard instructions that demonstrate intermediate pressure against falsification:

> In this blank you are asked for certain information about your background, your family, your home, your education, your hobbies, and your civilian employment. This is not a test. There are no "right" answers except those that tell the truth about yourself. To a large extent, your success in air-crew training will depend on how well you are understood by those in charge. All of the information asked for in this blank has been shown to be related to air-crew training. It is therefore to

your own interest to fill out this blank carefully and completely. You will record your answers on the separate answer sheet.

Read each question and the answers that follow it. Select the statement that best answers the question for you. Sometimes, no one of the answers will fit your case exactly. Do not worry about this, but select the answer that most nearly fits.

Approximately 600 to 900 students were tested at Psychological Research Unit No. 1 during May 1943 under each set of instructions with the results shown in table 27.7. Tables 27.8 and 27.9 show the critical ratios of the differences in means and in validity coefficients.

TABLE 27.7.—*Validity data for Biographical Data Blank, CE602D, under three different instructions based upon pilots in primary training and the graduation-elimination criterion*

Condition	N_t	p_g	M_g	M_e	M_t[1]	SD_t	r_{bis}
Minimum pressure for honesty	914	0.84	27.6	24.3	27.1	6.67	0.28
Maximum pressure for honesty	912	.80	25.9	21.2	25.0	6.73	.38
Standard instructions	661	.79	26.9	22.3	26.0	6.89	.40

[1] M_t=Mean of total group.

TABLE 27.8.—*Critical ratios of the differences between means of the total groups for Biographical Data Blank, CE602D, under three different instructions*

Condition	1	2	3
1. Minimum pressure for honesty	....	6.72	3.28
2. Maximum pressure for honesty	6.72	....	2.79
3. Standard instructions	3.28	2.79	....

TABLE 27.9.—*Critical ratios of the differences between validity coefficients for Biographical Data Blank, CE602D, administered under three different instructions*

Condition	1	2	3
1 Minimum pressure for honesty	....	2.40	2.66
2 Maximum pressure for honesty	2.40	....	.46
3 Standard instructions	2.66	.46	....

It can be seen that the biserial correlation for the Biographical Data Blank, CE602D, is highest (0.40) when administered under standard instructions. It was concluded that, although varied instructions do have apparent effects upon scores for this test, the differences are not much as to impair the effectiveness of the blank under standard conditions. Consequently, the standard instructions were retained in the administration of the test as part of the classification battery. It is most reassuring to note that the strongest possible pressure for honesty did not improve the validity of the test. Also noteworthy is the fact that encouragement of laxitity in responding seemingly lowered the test validity.

(4) *Scoring.*—Biographical Data Blank, CE602D, is scored in two ways—with a pilot key and a navigator key. The pilot key is based on item validities determined in a sample of 1,882 primary pilot graduates

and 735 eliminees (classes 42., 42J, 42K and 43A) tested by Psychological Research Unit No. 3 and also on 420 graduates and 176 eliminees tested by Psychological Research Unit No. 1 during February to April 1942. All cases were obtained from administration of Biographical Data Blank, CE602B. The development of this key is described in the description of the CE602B form. The navigator key was also based on this form, administered in January 1942 by Psychological Research Unit No. 1 to 312 examinees, including 270 graduates from navigation training and 42 eliminees. Both pilot and navigator keys include positively and negatively weighted responses. The number of responses receiving weights (either plus or minus) is 100 for pilot and 40 for navigator. The scoring formula in each case is R−W+20, where R refers to positively weighted and W to negatively weighted responses.

A special item-validity study.—A study was initiated at Psychological Research Unit No. 3 of the items in Biographical Data Blank, CE602B-SA and CE602D, to determine new weights for the items which might increase the predictive value of the stanine. Several hypotheses were guiding principles in this connection:

a. Items that are significantly related to graduation-elimination but unrelated to the stanine should make unique contributions to the stanine and so receive substantial weights.

b. Items that are significantly related to the stanine, but unrelated to graduation-elimination, should have negatively weights in the stanine, acting as suppression variables.

c. Certain items may add most to the stanine if weighted one way for cadets who have high stanines and another way for cadets who have low stanines.

Using answer sheets for Biographical Data Blank, CE602B-SA, completed by pilot students in class 43-H (tested at Psychological Research Unit No. 3), phi coefficients were computed for each alternative response for the following comparisons: (1) 660 graduates and 120 eliminees with high (7, 8, and 9) pilot stanines; (2) 148 graduates and 248 eliminees with low (1, 2, and 3) pilot stanines; (3) high-stanine and low-stanine graduates; and (4) high-stanine and low-stanine eliminees. Various combinations of these groups gave the correlation tables desired. The Biographical Data Blank was not in the classification battery at the time the pilots in this study were classified.

Signficant results appear when the items in the blank are considered in relation to graduation-elimination and to the stanine. The 76 responses that have significant phis with graduation-elimination, but zero phis or phis of opposite sign with the stanine, were weighted plus 1 or minus 1 to give the validity key for score A. The 92 responses that have significant phis with the stanine, but zero phis or phis of opposite sign with graduation-elimination, were weighted plus 1 or minus 1 to give the stanine key for score B.

A sample of all papers available was scored on these two keys, excluding the papers used in the original analysis.[6]

Table 27.10 shows the correlations with the stanine and with the graduation-elimination criterion of scores A and B.

TABLE 27.10.—*Validation data for Biographical Data Blank, CE602B-SA, for Keys Which Emphasize Correlation with the Criterion (Key A), and Correlation with the Stanine (Key B), Using Pilots in Primary Training and the Graduation-Elimination Criterion*

[$N_t=851$,[1] $p_t=0.61$]

Key	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]	r_{tet}	$_cr_{tet}$[2]	r_c[3]
A	20.76	18.24	4.50	0.35	0.30	−0.01	−0.03	0.09
B	19.06	19.31	3.00	−.05	.09	.12	.30	−.005

[1] Using only those with stanines of 4, 5, and 6, i. e., those not used in the item-validity study.
[2] Assuming an unrestricted stanine standard deviation of 2.00.
[3] The amount of validity that the score would add to the validity of the pilot stanine if the test were added to the Classification Battery of August 1942.

It can be seen that score A would make a significant addition to the stanine in use for these students, but score B would not. The failure of score B to add anything is partly a result, however, of the selection of items. Items having significant relationships both to the stanine and to graduation-elimination, but of opposite sign, logically might have been included in either score. To avoid duplication, these items were weighted only in score A. If, instead, they had been weighted in score B, this score would probably have approached a significant addition to the stanine.

The correlation between scores A and B is −0.14, both raw and corrected. Thus, though the validity of score B is low, it is measuring something valid not included in score A. A positive correlation between the two scores would have been a more favorable condition for the use of one as a suppression variable.

An examination was made of the responses in CE602B-SA that discriminate significantly in one direction between graduates and eliminees at one stanine level but fail to discriminate or discriminate in an opposite direction at another stanine level. No consistent trend appeared in the analysis of the 25 items in this category. The items come from all sections of the blank. The only general finding in part I of the blank (factual items) appears to be a slight indication that the low-stanine graduates are men who have not had experience in certain areas (such as mechanical) which are heavily weighted in the stanine, but who may have been potentially able in those areas. The opinion items that discriminate differentially also seem to represent no significant pattern. In either case, the number of items manifesting this type of discrimination is so small that chance could have produced the relationships in question.

In conclusion, this study showed that it seems doubtful that biographical items discriminate differentially between graduates and eliminees at

[6] The only cases unused in the item analyses were those having pilot stanines of 4, 5, and 6.

different stanine levels to a significant degree. Items that are related to graduation-elimination, but unrelated or oppositely related to the stanine, add very significantly to the predictive value of the stanine. Items that are related to the stanine but unrelated to graduation-elimination do not add to the stanine prediction.

Since a valid key independent of the stanine had been prepared successfully, it was decided to try the same procedure for form CE602D. Utilizing 600 graduates and 600 eliminees (classes 44–D through 44–H) from primary training, who had been tested with the July 1943 battery, two keys were again prepared, one to maximize validity, and one to maximize correlation with stanine. A new sample of 2,000 cases from classes 44–I and 44–J (July and November 1943 batteries) was utilized for validation of these keys. The results are shown in table 27.11.

TABLE 27.11.—*Validation data for Biographical Data Blank, CE602D, for the classification battery key and for keys which emphasize correlation with the criterion (key A) and correlation with the stanine (key B), using pilots in primary training and the graduation-elimination criterion* [N_t=2,000, p_e=0.86]

Key	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]	r_{sts}	$_cr_{sts}$[2]	r_s[3]
Classification	30.10	27.28	6.27	0.24	0.32	0.46	0.55	....
A	26.14	24.48	4.76	.22	.26	.27	.33	0.015
B	24.44	23.32	4.58	.12	.18	.34	.41	.000

[1] In classes 44I and 44J.
[2] Assuming an unrestricted stanine standard deviation of 1.83.
[3] The amount of validity the score would add to the validity of the pilot stanine.

It is apparent from these data that the procedure of selecting items to correlate with the stanine and not with the criterion, or vice versa, was less successful with form CE602B and the July and November stanines than with form CE602B–SA and the August stanine. This is due in part to increased correlation with blank and stanine. It is also due in part to the fact that the biographical-data score was included in the stanine used as a criterion for item analysis. The more accurate procedure of subtracting the biographical-data score from the stanine before doing the item analyses and correlations with stanine did not seem worth the considerable labor entailed at the time the study was undertaken. It was believed that the unique contribution of Biographical Data to stanine validity was larger than is now known to be true.

In summary statement we may say: (*a*) The procedure used in these studies gave promising results with the first blank and stanine, in that one group of items related to graduation-elimination but unrelated to the stanine added a significant amount to the validity of the stanine; (*b*) when this procedure was tested with a later version of the blank, the amount of additional validity decreased markedly; and (*c*) significant negative contributions to stanine, using keys which maximized correlation with stanine and minimized correlation with criterion, were not found for either blank.

Statistical results. (1) *Distribution statistics.*—Score distributions for form D are described by the data in table 27.12.

TABLE 27.12.—*Distribution statistics for pilot and navigator scores of Biographical Data, CE602D*

Score	Group	N	M	SD
Pilot	Unclassified Aviation Students[1]	3,000	27.5	6.9
Do	....do[2]	1,920	26.8	6.5
Do	....do[3]	1,500	27.2	6.8
Do	West Point cadets[4]	888	28.7	6.9
Navigator	Unclassified Aviation Students[1]	3,000	22.6	3.2
Do	....do[2]	1,920	21.8	3.0
Do	....do[3]	1,500	22.3	3.0
Do	West Point cadets[4]	888	24.8	2.8

[1] Tested at Psychological Research Units Nos. 1, 2, and 3 with the July 1943 Classification Battery.
[2] Tested at Medical and Psychological Examining Units Nos. 4 to 10 inclusive with the November 1943 Classification Battery.
[3] Tested at Psychological Research Units Nos. 1, 2, and 3 with the November 1943 Classification Battery.
[4] Class of 1946.

(2) *Reliability coefficients.*—A reliability coefficient of 0.62 (corrected) was obtained for the pilot score with a sample of 1,000 unclassified students tested at Medical and Psychological Examining Unit No. 7. For the navigator score on the same sample, the coefficient was 0.35 (corrected). These coefficients were obtained by the split-half method. An attempt was made to choose items for each part that would make the content of the halves roughly comparable. The coefficients obtained are much lower than test-retest coefficients, which are 0.86 and 0.49 respectively, for a time interval of approximately 28 days (N=711, on classified aviation students tested at Medical and Psychological Examining Unit No. 6, 11 to 19 April 1945).

(3) *Factor composition.*—The classification form of the Biographical Data Blank, CE602D, was factor analyzed with two batteries (see ch. 28 for a full discussion), the July 1943 Classification Battery and the November 1943 Classification Battery, and for both the navigator and pilot keys.

The navigator score in both analyses revealed only one significant loading, which is in the mathematics-background factor. The loading in the July 1943 battery was 0.42, in the November 1943 battery 0.50, with a weighted average of 0.45. The communality is quite low, as might be expected from the dearth of significant loadings in diverse factors. In the July battery the communality was 0.25, in the November battery, 0.30, with a weighted average of 0.26.

For the pilot score of the blank the highest loading is in the mechanical-experience factor. This loading was 0.50 in the July-battery analysis and 0.53 in the November-battery analysis, with a weighted average of 0.51. The next highest loading appeared in the mathematics-background factor, which for the July battery was 0.29, and for the November battery 0.31, with a weighted average of 0.30. A consistent negative load-

TABLE 27.13.—*Validity data for the pilot score on Biographical Data Blank, CE602D*

Group	Class	Psychological Research Unit No.	N_t	P_g	M_g	M_e	SD_t	r_{bis}	$_sr_{bis}$[1]
Pilots in primary training	44D, 44E	1	1,950	0.87	27.70	24.50	6.70	0.26	[2]0.35
Do	44E	1, 2, 3	4,779	.88	28.55	25.06	6.38	.29	[3].32
Do	44I	1, 2, 3	3,145	.84	30.16	28.47	5.97	.16	[4].26
Do	44J	([5])	1,676	.89	30.05	28.60	6.16	.12	[2].26
Pilots through advanced training	44E	1	2,344	.74	28.47	24.91	6.67	.32	[4].38
Do	44E	2	1,822	.80	29.03	25.60	6.49	.30	[4].33
Do	44E	3	2,955	.80	28.77	26.59	6.18	.20	[4].26
Navigators	([5])	2, 3	1,[illegible]49	.81	25.75	24.53	6.79	.10	[2].18
WASPs	44-W-8	([5])	104	.61	22.57	19.42	5.97	.33	....
Air mechanics in training	([5])	[7]6	428	....	....	....	....	[8]−.01	....
Do	([5])	[7]6	240	....	....	....	....	[8]−.19	....
Armament trainees	34-44A, 35-44A	[7]1-10	269	....	....	....	....	[8]−.08	....

[1] All biserial r's are for the graduation-elimination criterion. For the samples followed through advanced training, this criterion includes an unknown number of eliminations for fear or own request.
[2] Assumed unrestricted stanine standard deviation not identified.
[3] Assuming an unrestricted stanine standard deviation of 2.00.
[4] Assuming an unrestricted stanine standard deviation of 1.90.
[5] Not identified.
[6] Assuming an unrestricted stanine standard deviation of 1.80.
[7] Medical and Psychological Examining Units.
[8] Product-moment correlations, using final average grades as criteria.

TABLE 27.14.—*Validity data for the navigator score on Biographical Data Blank, CE602D*

Group	Class	Psychological Research Unit No.	N_t	r_t	M_g	M_e	SD_t	r_{bis}[1]	$_c r_{bis}$
Navigators	(2)	2, 3	1,249	0.81	23.97	23.33	3.09	.12	[3]0.34
Pilots in primary training	44P	1, 2, 3	4,779	.88	22.68	22.63	3.10	.01	[4].01
Do	44I	1, 2, 3	3,145	.84	22.28	22.24	2.94	.01	[4].02
Do	44J	....	1,676	.89	21.83	21.61	2.93	.04	[5].05
Pilots through advanced training	44E	1	2,346	.74	22.69	22.68	3.04	.00	[6].01
Do	44E	2	1,823	.80	22.98	22.77	3.13	.04	[6].04
Do	44E	3	2,995	.80	22.48	22.19	3.18	.05	[6].06
WASPs	44-W-8	(2)	104	.61	21.59	21.05	2.87	.12	....
Air mechanics in training	(2)	[7]6	428	....	....	....	....	[8].00	....
Do	(2)	[7]6	240	....	....	....	....	[8].01	....
Armament trainees	34-44A, 35-44A	[7]1-10	269	....	....	....	....	[8].03	....

[1] All biserial r's are for the graduation-elimination criterion. In the samples followed through advanced training, this criterion includes an unknown number of eliminations for fear or own request.
[2] Not identified.
[3] Assumed unrestricted stanine standard deviation not identified.
[4] Assuming an unrestricted stanine standard deviation of 2.00.
[5] Assuming an unrestricted stanine standard deviation of 1.90.
[6] Assuming an unrestricted stanine standard deviation of 1.80.
[7] Medical and Psychological Examining Units.
[8] Product-moment correlations, using final average grades as criteria.

ing in the numerical factor appeared: −0.20 for the July battery, −0.26 for the November battery, with a weighted average of −0.22. The communality for the pilot score is higher than for the navigator score, being 0.46 for the July battery, 0.55 for the November battery, and 0.48 for the weighted average.

(4) *Test validity.*—For validity of the pilot score see table 27.13, and for the navigator score, table 27.14.

Evaluation.—This form of the Biographical Data Blank (CE602D) was refined to the point where it was acceptable for the classification battery. All nonpredictive items had been dropped, leaving only 65. The data show that this form is a good predictor of both pilot and navigator aptitude, and, as such, is a useful instrument in classification testing.

Both scores are valuable because of their unique elements. Mathematical background makes up about 18 percent of the total variance of the navigator score—more than the test mathematics A has to offer toward this factor. This score should be made more reliable by increasing the number of keyed responses, at the same time increasing the variance in mathematics background and perhaps adding other valid variance.

About 26 percent of the variance in the pilot score is in mechanical experience, which is covered better by mechanical tests. Its unique validity is due to an unknown factor or factors. This unknown variance should be identified and enlarged. For the sake of purity the mechanical variance should be removed. The variance in mathematical background should also be removed from the pilot score. To it can be attributed the small validity of the pilot score for navigation training. The purposes of classification would be better served if this variance were confined to the navigator score.

Variations. (1) *Biographical Data Blank, CE602C.*—As mentioned above, this form is made up of 65 items which showed empirical validity for pilot or navigator prediction. Since it is exactly like CE602D except for order of items, nothing further need be said concerning it.

(2) *Form CE602W.*—A special form of the Biographical Data Blank, CE602W, was constructed at Headquarters, AAF Training Command, for administration to WASP (Women's Auxiliary Service Pilots) pilot trainees.[a] While no statistical data are available at the time of writing, a brief description of the form may be of interest. This blank contained 61 items, which were slanted for a female population and designed to elicit general biographical information, expressions of interest, and facts of experience. The categories covered include age, history of minor illness, flying experience, marital and maternal status or plans, social life in college, marital adjustment of parents, use of cosmetics, tobacco, and liquor, childhood and adult activities and interests. Some items were suggested by the impressions gained by a psychiatrist while interviewing a preliminary group. He found that a number of health and per-

[a] Chief contributor: Maj. R. L. Thorndike.

sonal factors, particularly relating to marriage, seemed significant in the general adjustment of these trainees. Following are several typical items:

Do you have headaches?
- A. Practically never have headaches.
- B. Have mild headaches occasionally at irregular intervals.
- C. Have occasional severe headaches.
- D. Have headaches recurring at regular intervals.
- E. Have frequently recurring severe headaches.

If you are married or get married, what would you consider the ideal size of family for you?
- A. No children.
- B. One child.
- C. Two children.
- D. Three or four children.
- E. Five or more children.

While you were in college (or of college age), how frequently did you have "dates?"
- A. Not at all.
- B. Once or twice a month.
- C. About once a week.
- D. Two or three times a week.
- E. Four times a week or more.

Biographical Data Blank, CE602E[10]

In an attempt to increase the validities of form CE602D and to provide a more adequate coverage of personality and background factors related to air-crew training, new items were constructed and incorporated in CE602E. These new items were suggested by inspection of (1) valid items in the CE602D version of the blank, (2) job-analysis data, (3) clinical data, and (4) experimental personality inventories.

Description.—Part I of the blank pertains to social background in general, and part II pertains to self-evaluations of different personality aspects. Categories of items in part I include: early home environment, hobbies, athletic experience, career and occupational interests, education of parents, father's occupation and financial condition, intrafamily relationships, extent of parental participation in examinee's home and school life, marital interests, social habits and types of friends, personality traits considered desirable, sleep habits and dreams, opinions about Army officers, and flying-duty preferences.

Items in part II pertain to the following categories: vacation preferences; leisure-time activities and satisfactions; interests in books, songs, and comics; opinions about typical social-conflict situations; self evaluation of personality traits; opinions about behavior of other people; opinions about the enemy; and opinions about generally accepted social attitudes.

(1) *Internal characteristics.*—This form contains 300 items, divided into 2 parts of 150 each.

[10] Developed at Psychological Research Unit No. 1. Chief contributors: Lt. John S. Harding, Sgt. Harold M. Proshansky, Sgt. [illegible] Srole, and Capt. Donald E. Super.

(2) *Administration.*—All examinees are supposed to finish both parts of the blank. This requires approximately 60 minutes per part. Examinees are informed when half the time is up on each blank.

Following are two typical items from each part, respectively:

Which of the following skilled jobs would you rather hold?

A. Draftsman.
B. Carpenter.
C. Watchmaker.
D. Inspector.
E. Tool and die maker.

How often do you worry?

A. Very frequently.
B. Frequently.
C. Occasionally.
D. Very seldom.
E. Never.

The principal satisfaction most people get out of participating in sports is that of:

A. Showing their skills.
B. Being with friends.
C. Preserving health.
D. Competing with others.

Ordinarily, labor does not get its fair share of what it produces.

A. Yes.
B. No.

(3) *Scoring.*—It is planned that new valid items from test CE602E will be incorporated with items in CE602D. The scoring formula will be R−W+2.

Statistical results. (1) *Item validity (pilot).*—Form E was administered at Psychological Research Unit No. 1 in April 1944 to 682 classified pilots in preflight school, and the sample was divided into odds and evens groups for item correlation with the graduation-elimination criterion (primary training) and cross-validation of total scores. The proportion of graduates was 0.78. Responses with phis of 0.11 or greater were keyed. The distribution of phis is given in table 27.15.

TABLE 27.15.—*Frequency distribution of item-validity phis of responses in Biographical Data, CE602E, for pilot students*

Phi range	Odds sample[1]				Evens sample[2]			
	2-choice		Multiple-choice		2-choice		Multiple-choice	
	Part I	Part II	Part I	Part II	Part I	Part II	Part I	Part II
0.23 to 0.27 ...	0	1	2	0	0	0	0	0
.18 to .22 ...	1	0	7	0	0	4	0	0
.13 to .17 ...	3	6	12	5	3	3	15	1
.08 to .12 ...	7	15	42	15	4	27	36	14
.03 to .07 ...	7	54	68	18	17	51	91	25
−.02 to .02 ...	12	38	109	26	9	28	95	25
−.07 to −.03 .	....	....	68	26	....	....	79	27
−.12 to −.08 .	....	....	41	12	....	....	32	10
−.17 to −.13 .	....	....	8	2	....	....	3	2
−.22 to −.18 .	....	....	6	0	....	....	2	0
Totals ...	30	114	363	104	33	113	353	104

[1] N=340.
[2] N=342.

No items with more extreme division than 85—15 were included. Items having two alternative responses were segregated from those having more than two, and only positive phis are given in them.

(2) *Test validity (pilot)*.—Two scoring keys were derived from the odds and evens, and a cross-validation study was made. The data are found in table 27.16.

TABLE 27.16.—*Validity data for two empirical keys for Biographical Data, CE602E, pilot score (p_e=0.78)*

Groups	Key	Part	Formula	N_t	M_g	M_e	SD_t	r_{bis}	$_cr_{bis}$[1]
Evens .	Odds[2]	I	Rights ...	342	13.75	12.68	3.54	0.18	0.22
Do	....do[2]	I	Wrongs .	342	9.18	9.78	2.76	−.13	−.13
Do	....do[2]	I	R−W ...	342	4.57	2.90	5.78	.17	.20
Odds ..	Evens[3]	I	Rights ...	340	12.03	10.99	2.77	.22	.21
Do	....do[3]	I	Wrongs .	340	4.56	5.20	1.69	−.22	−.24
Do	....do[5] .	I	R−W ...	340	7.47	5.79	3.86	.25	.26
Evens .	Odds[4]	II	Rights ...	342	8.27	8.00	2.03	.00	−.01
Do	....do[4]	II	Wrongs .	342	6.20	6.38	1.78	−.06	−.04
Do	....do[4]	II	R−W ...	342	2.07	1.62	3.52	.03	.01
Odds ..	Evens[5]	II	Rights ...	340	9.16	9.04	1.96	.03	.04
Do	....do[5]	II	Wrongs .	340	7.71	7.76	1.88	−.02	−.02
Do	....do[5]	II	R−W ...	340	1.45	1.28	3.67	.03	.03

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Number of scored items=59.
[3] Number of scored items=41.
[4] Number of scored items=34.
[5] Number of scored items=36.

(3) *Item-validity (navigator)*.—Form E was administered to 897 (part I) and 888 (part II) classified navigation students. Testing was accomplished at Ellington Field and Selman Field in February 1944 and at Psychological Research Unit No. 3 in February and March 1944. Each group of students was subdivided into odds and evens samples, and item correlations with the graduation-elimination criterion (advanced training) were computed. The proportion of graduates was 0.89. Responses with phis of 0.10 or greater were keyed. The distribution of phis are given in table 27.17.

TABLE 27.17.—*Frequency distributions of item-validity phis of responses for Biographical Data Blank, CE602E, for navigation students*

Phi range	Odds sample				Evens sample			
	2-choice		Multiple-choice		2-choice		Multiple-choice	
	Part I	Part II	Part	Part II	Part I	Part II	Part I	Part II
0.23 to 0.27 .	1	0	2	0	1	2	1	0
.18 to .22 ..	1	3	5	1	1	1	10	2
.13 to .17 ..	2	9	11	6	5	6	21	8
.08 to .12 ..	4	26	52	9	4	37	43	20
.03 to .07 ..	22	63	82	21	12	45	90	17
—.02 to .02 ..	8	22	101	19	8	21	108	27
—.07 to —.03 .	....	....	88	18	....	....	88	20
—.12 to —.08	....	....	42	11	....	....	37	17
—.17 to —.13 .	....	....	17	5	....	....	23	6
—.22 to —.18 .	....	....	3	3	....	....	8	1
—.27 to —.23	....	....	0	0	....	....	3	0
Totals ...	38	123	403	93	31	112	432	118

(4) *Test validity (navigator)*.—Two scoring keys derived from odds and evens groups were used in the cross-validation study. The results are given in table 27.18.

703320—47—51

TABLE 27.18.—*Validity data for two empirical keys for Biographical Data, CE602F, navigator score*

[$p_s = 0.89$]

Group	Key	Part	Formula	N_s	M_e	M_s	SD_s	r_{bis}	$_c r_{bis}$[1]
Evens ..	Odds[2]	I	Rights ...	449	21.41	20.88	1.49	0.19	0.25
Do	....do[2]	I	Wrongs ..	449	13.91	14.57	2.89	−.12	−.21
Do	....do[2]	I	R−W ...	449	7.50	6.31	4.33	.14	.23
Odds ..	Evens[3]	I	Rights ...	448	22.67	21.98	3.48	.10	.10
Do	....do[3]	I	Wrongs ..	448	19.18	19.41	3.20	−.04	−.10
Do	....do[3]	I	R−W ...	448	3.49	2.57	5.80	.08	.11
Evens ..	Odds[4]	II	Rights ...	442	18.01	18.24	3.18	−.04	.05
Do	....do[4]	II	Wrongs ..	442	15.88	15.51	2.70	.07	.10
Do	....do[4]	II	R−W ...	442	2.13	2.73	5.57	−.06	.02
Odds ..	Evens[5]	II	Rights ...	446	28.08	28.21	3.82	−.02	−.02
Do	....do[5]	II	Wrongs ..	446	20.03	19.30	3.22	.12	.14
Do	....do[5]	II	R−W ...	446	8.05	8.91	6.35	−.07	−.08

[1] Assuming an unrestricted stanine standard deviation of 2.00.
[2] Number of scored items=61.
[3] Number of scored items=65.
[4] Number of scored items=41.
[5] Number of scored items=62.

Evaluation.—Part I of the inventory proved to be moderately valid for the selection cf pilots. Its correlation with the stanine was so low (even though the stanine includes as one component the score on Biographical Data Blank, CE602D) that considerable uniqueness is evident. The new items in part I will therefore add noticeably to the pilot validity of the classification form of the Biographical Data Blank. Part I promises some additional validity as scored for navigator selection.

Part II does not show any promise of validity for either specialty, even though there appear to be a few valid items. It is worthy of comment that part II is more devoted to questions concerning the examinee's personality traits and less to the factual type of part I.

Biographical Data and Pilot Specialization

Two forms of this type of test were developed especially with the view of discriminating between promising fighter pilots and bomber pilots. Neither had been followed through at the time this chapter was written, but they will be described for the record.

(1) *Biographical Data Blank, CE602FW.*—This form was constructed at headquarters, AAF Training Command, Fort Worth.[11] It was administered to pilots in basic schools during October 1943 along with seven other experimental tests.

Although this study was concerned primarily with pilots, the possible prediction of navigator aptitude was considered important. Many instructors felt that navigational ability was important to a pilot in flying a heavy bomber. Thus, it was believed that information which would predict navigator success could help differentiate the bomber from the fighter pilot.

Most of the items in the Fort Worth Biographical Data Blank are identical with or similar to items in the previous forms of the Blank. This form contains 147 items divided into the usual categories.

[11] Chief contributor: Capt. Launor F. Carter.

(2) *Biographical Data Blank, CE602F.*[12]—There were two main criteria for the selection of items in CE602F. First, no questions were to be taken from the classification-battery form of the Blank, CE602D. Second, no questions were to be taken from the Fort Worth form, which had been administered previously in a pilot specialization project. Consequently, items that were deemed likely to discriminate between fighter and bomber pilots either were selected from other forms of the Biographical Data Blank or were written originally for CE602F. The categories include marital and parental status, experience in saving money, entertainment preferences, study habits, summer-camp experience, relationships to parents, and home environment.

Test CE602F contains 61 items divided into the categories described above, and requires approximately 20 minutes for administration.

Following are two typical items:

To how many social clubs or organizations have you belonged?
- A. None.
- B. 1.
- C. 2.
- D. 3.
- E. 4 or more.

When you have a little extra money, you usually:
- A. Buy some luxury you have wanted for a long time.
- B. Get a good meal in town.
- C. Go on a date.
- D. Go out with the "boys."
- E. Save it.

Occupational Experience Blank, CE603A [13]

This blank was designed to reveal information from which examinees could be classified occupationally according to training, interest, and experience. It was based on the hypothesis that different occupational groups possess different average air-crew aptitudes, and that prediction of success or failure in air crew can be improved by knowledge of the examinee's previous work experience.

Description.—This test asks for specific information concerning occupations and training of students before and during their Army careers.

(1) *Internal characteristics.*—The blank is divided into five parts. Part I asks for information as to the subjects studied in high school, college, and vocational school, and the number of semesters they were studied. Part II asks for information concerning special skills or abilities acquired outside the work experience. Part III provides for a detailed description of full-time civilian jobs and a briefer description of part-time and temporary jobs. Part IV provided for a description of training and duty assignments in the Army. Part V provided for an ex-

[12] Developed at Psychological Research Unit No. 3. Chief contributors: Cpl. Stanley Blumberg, Lt. John I. Lacey, and Lt. Eli A. Lipman.
[13] Developed at Psychological Research Unit No. 1. Chief contributor: Capt. Seymour P. Stein.

pression by the student of his occupational interests. There is a one-page supplement to the blank which asks eight questions of biographical information.

(2) *Administration.*—All examinees are supposed to finish the blank. The administrator takes up one item at a time, explaining it thoroughly before the examinees answer the item. The time required is 45 minutes.

(3) *Scoring.*—Scoring consists of classifying the blanks according to the Dictionary of Occupational Titles of the United States Employment Service. The intention was to obtain quite specific classifications and not to lump various occupations into broad categories. After the examinees were classified, they were divided into occupationally homogeneous groups. After primary training, each occupational group was to be compared to the whole group to see whether there is a significant difference in the graduation rate.

Statistical results.—No data are available.

Personal Data form, CE605A

The purpose of this inventory [14] is to measure susceptibility to combat and near-combat neuroses by means of carefully selected biographical items.

Description.—The problem of predicting combat neurosis, or susceptibility to neurosis, is one which was bypassed during the early days of the AAF Psychology Program due to the emphasis on rapid development of selective instruments that validated against the criterion of successful completion of some type of air-crew training. As selective techniques became more refined, and the war progressed to the extent that combat criteria were becoming available, some attention swung to the problem of development of instruments that might predict successful combat performance and susceptibility to combat fatigue or combat neurosis. Several projective procedures were developed with this in mind (see ch. 24). This instrument, however, represents the first biographical-data approach to the problem. The rationale underlying the test rests on the assumption that an individual's history, to the extent that it can be obtained and correctly evaluated, is the best single index of his future performance.

(1) *Internal characteristics.*—This instrument consists of 139 items, each with from 2 to 5 alternative responses. These items deal primarily with aspects of familial status and personality development. Some of the items were taken from existing biographical inventories, and others were written especially for this test. Sample items are:

As a child the teachers I liked best were:

A. Middle aged women.
B. Young women.

[14] Developed at Psychological Research Unit No. 1. Chief contributors: Sgt. Gerald S. Blum, Lt. Vivian Fisher, and Capt. Donald E. Super.

C. Middle aged men.
D. Young men.

My parents always considered my behavior:
A. Much better than that of other children.
B. Slightly better than that of other children.
C. Slightly worse than that of other children.
D. Much worse than that of other children.

(2) *Administration.*—Pertinent directions are:

In this booklet you are asked for certain information about your personal history. This is not a test in the ordinary sense; there are no "right" answers except the ones which reflect your own particular past experiences and situations.

To a large extent your success in flying depends on how well you are understood by those in charge. All of the information asked for in this booklet is for the purpose of aiding your superior officers in understanding you. It is to your own advantage, therefore, to indicate your answers to the items in this booklet as carefully, completely and honestly as you can.

Whenever the word "parents" or "father" or "mother" is used in the following questions and statements, it will be understood to include, when appropriate or fitting in your case, any such words as "foster parents," "adopted parents," "step parents," "legal guardians" or "foster father" * * *

(3) *Scoring.*—A priori keys were not used. Scoring was accomplished and validities were obtained by means of cross-validation data using a training criterion. Two scores were obtained; one based on positively weighted responses and one on negatively weighted responses.

Statistical results.—The data that follow were computed for a sample of 738 pilots in primary training, originally tested in May 1944 at Psychological Research Unit No. 1.

(1) *Item validity.*—The distribution of phis based on item analysis used in the cross-validation study is presented in table 27.19.

TABLE 27.19.—*Distribution of phis based on a sample of 738 pilots in primary training, using a graduation-elimination criterion, for the Personal Data Form, CE605A*

Phi	f (odds)	f (evens)	Phi	f (odds)	f (evens)
0.18 to 0.21...	2	2	—.07 to —.03 ..	85	84
.13 to .17...	13	8	—.12 to —.08 ..	49	37
.08 to .12...	35	44	—.17 to —.13 ..	10	14
.03 to .07...	85	88	—.22 to —.18 ..	3	1
—.02 to .02...	106	111			

In order not to confuse the form of the phi distribution, the three questions of a two-choice form which are contained in the test were dropped from this listing. None of these reached or exceeded the 5 percent level of significance.

(2) *Test validity.*—Validation data were computed for a sample of 738 pilots, which was split into two equal groups, odds and evens. Separate item analyses were accomplished for each subsample, and two scoring keys devised. The criteria for scoring a response were: (a) a phi

significant at or beyond the 5 percent level (0.10); (b) a split of 18–15 or better.

The evens group was scored with the odds key, and the odds group was scored with the evens key. The validities obtained are presented in table 27.20.

Table 27.20.—*Validity data based on two groups of pilots in primary training, using a graduation-elimination criterion, for the Personal Data Form, CE605A*[1]

Group	Score	M_p	M_e	SD_t	r_{bis}	$_cr_{bis}$[2]
Odds scored with evens key.[3]	Rights	11.21	10.78	2.78	0.09	0.09
	Wrongs	8.91	9.29	2.38	−.09	−.08
Evens scored with odds key.[4]	Rights	12.26	11.33	2.74	.09	.10
	Wrongs	13.23	13.17	2.93	.01	−.05

[1] For both groups N_t=369, p_p=0.81.
[2] Corrected to an unrestricted stanine standard deviation of 2.00.
[3] Number of scored items=38.
[4] Number of scored items=46.

For an N of 369, an r_{bis} of 0.10 is significant at the 5 percent level of confidence, and an r_{bis} of 0.13 is significant at the 1 percent level. It can be seen that none of the uncorrected correlations were significant at either of these levels.

Evaluation.—There would seem to be an excessive number of phis beyond the confidence limits, but, in view of the apparent unimodality of the distribution with its central tendency at zero, and the failure of the cross validation test to show significant biserial correlations, it is probable that there are few, if any, genuinely valid items in this collection for the prediction of primary pilot training success.

While a relatively large number of individual items appeared to be valid for the prediction of success in primary pilot training, the biserial correlations for total scores did not support that promise of validity. It is to be remembered that this instrument was designed to predict susceptibility to combat neurosis and that combat criteria were not employed in obtaining test validity. A validation study that will test the original hypothesis concerning the value of this test is still to be made.

EVALUATION AND CONCLUSIONS

Statistical results for some of the tests reviewed in this chapter reveal the empirical truth of the hypothesis that aptitude for air-crew training can be predicted from certain biographical information. The classification-battery form of the Biographical Data Blank (CE602D) was shown to be a satisfactory measure predictive of pilot and navigator success.

No attempt was made to develop a scoring key for the bombardier. The reasons were several. The task of the bombardier, unlike those of pilot and navigator was without precedent either in military or civilian life, and is, therefore, somewhat characterless. Its resemblances to vocations or avocations are limited, and so hypotheses regarding items are difficult to invent. Another reason was the lack of a training criterion in which reasonable confidence could be placed. Circular error was highly unreliable as a measure of bombing accuracy; and graduation, which depended heavily upon it, was hard to predict by the best of tests. Coupled with this was the fact that numbers of trainees in the early days were small, and validation data in sufficient quantities were slow in accumulating. Failure to develop a valid key for the bombardiers in the General Information test was also discouraging of success with the Biographical Data Blank. For completeness in a research program, however, the validation of items for bombardier would have been desirable.

Factor analyses did not reveal all the reasons why the Biographical Data Blank is valid for pilots. The test has considerable unknown valid variance. It is toward the understanding of this unique variance and its further exploitation that new work should be directed in order to improve the pilot score. While the valid variance of the navigator score is probably fully accounted for, this score could be considerably improved in reliability and therefore in validity. This would mean the search for a large number of new items emphasizing the mathematical-background factor, in order to maximize the usefulness for the prediction of navigation success. When a satisfactory criterion is found for the bombardier, attention should be given to the writing and validation of items for that specialty.

BIBLIOGRAPHY

(1) Guthrie, Edwin R. "Personality in Terms of Associative Learning" in *Personality and Behavior Disorders*, ed. by Hunt, Joseph McV., New York: The Ronald Press Co., 1944, Vol. I.

(2) Johnson, H. M., in cooperation with Boots, Mary L. and Wherry, Robert J. *On the Actual and Potential Value of Biographical Information as a Means of Predicting Success in Aeronautical Training*, Washington, D. C.: Civil Aeronautics Administration, August 1944.

(3) Kelly, E. L. and Ewart, E. A preliminary study of certain predictions of success in civilian pilot training, Washington, D. C.: Civil Aeronautics Administration, December 1942.

CHAPTER TWENTY-EIGHT

Factorial Picture of Tests and Criteria[1]

INTRODUCTION

Reasons for a Factorial Picture

It is the purpose of this chapter to summarize and to evaluate what is known concerning the factorial composition of the tests described in this volume, and of the criteria that they were designed to predict. This is done with the conviction that the most significant information one can have concerning either tests or criteria is in the form of factorial description. The common factors serve as joint reference categories for both alike.

Factorial knowledge of tests enables us to predict validities in advance if we also know the relative weights of the factors in the criteria. Generalizations can be made, therefore, beyond the usual facts regarding validities of specific tests for specific criteria. Having a large battery of tests covering most of the human traits, each test described in terms of factors and their loadings, one could then readily fit tests to new selective uses merely by analyzing the new criteria. Having the specifications for a job reduced to terms of factor loadings, one would then be ready to select a battery of tests which would yield near maximum validity. If pure tests of factors have been constructed, an economical battery is assured. Pure tests can be arrived at by factorial procedures.

Plan of the Chapter

Several factorial studies have been reported in preceding chapters, each where it most appropriately applied to a group of tests. Before presenting an over-all summary, it will be necessary to give an account of four general analyses whose results enter prominently into consideration. These analyses were based upon four of the classification batteries —those of December 1942, July 1943, November 1943, and September 1944. These analyses will be presented and then a list of the common factors with the best available definitions. Tests will then be grouped according to their leading factors and tabulated with weighted-average factor loadings, with communalities based upon these averages, with estimates of reliability, and with weighted-average validities for pilot, bombardier, and navigator training where those data are available. A

[1] Written by the Editor.

list of tests in alphabetical order with the same kind of data will be given in Appendix B.

One feature of the chapter is almost unique in vocational psychology. There are presented estimated factor loadings for the three training criteria, as was forecast at the end of chapter 1. From the estimates of loadings of factors not now represented in the classification battery, we can see just what types of tests might well be added to the battery in order to increase the coverage of the criteria and so improve predictions of success.

ANALYSIS OF CLASSIFICATION BATTERIES

The Data

It was a general policy to obtain intercorrelations of tests in all classification batteries very early after they went into effect, based on very large samples of unclassified aviation students from different examining units. These intercorrelations were not only used later as a basis for revising regression weights, but were also good material for factorial studies. The analyses have been based upon unusually large samples, consequently, the results should be quite stable.

The December 1942 Battery.[2]—For the December 1942 battery, the numbers of cases were 3,254 and 4,774, the smaller number applying to the new revised forms of tests going into the battery for the first time. To the matrix were added validity coefficients for the criteria for bombardier, navigator, and pilot training. Those for the bombardier were based upon 1,829 students (1,453 graduates and 376 eliminees). Those for the navigator were based upon z-averages of two samples, 1,970 (1,554 graduates and 416 eliminees) in one, and 731 (633 graduates and 98 eliminees) in the other. Those for the pilot were based mostly on z-averages from a total of 10,925 students. The intercorrelations of the criteria were guessed from previous estimates (see table 28.14) of their common-factor loadings. The correlation matrix for this battery is presented in table 28.1.

The July 1943 Battery.[3]—The intercorrelations for this battery are presented in table 28.2. They were based upon 3,000 unclassified students, 1,000 from each of the three original examining units. Along with this battery were included in the analysis four composite scores—a weighted aggregate for each air-crew specialty and the officer-quality composite score. The purpose of these inclusions was to determine how these composites were weighted factorially. (A more satisfactory method of estimating factor loadings in the stanines would have been by the correlation of weighted sums with each factor.) Being very complex, and having very high communalities, they made a factorial solution more difficult. No corrections were made for the spuriousness of the

[2] The analysis was executed by Capt. Lloyd G. Humphreys and Capt. John I. Lacey.
[3] The analysis was executed by S/Sgt. J. Gordon Ethin and Capt. Lloyd G. Humphreys.

TABLE 28.1.—*Correlation matrix for the December 1942 classification battery*
(N=3,254 to 4,774)[1]

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1. Technical Vocabulary—Pilot[2]	...	37	41	40	40	28	14	08	−06	−03	09	19	14	22	10	14	13	08	05	04	00	11	21
2. Technical Vocabulary—Bombardier	37	...	40	39	20	23	25	18	04	09	15	09	14	18	13	11	06	09	04	05	04	15	12
3. Technical Vocabulary—Navigator	41	40	...	56	15	29	58	38	15	22	28	11	16	22	16	12	06	05	03	07	05	33	09
4. Reading Comprehension	40	39	56	...	42	50	43	43	06	16	28	12	13	26	24	22	20	08	04	08	13	38	20
5. Mechanical Information	40	20	15	42	...	55	04	11	−17	−09	00	08	03	16	05	22	34	10	04	06	00	08	28
6. Mechanical Principles	28	23	29	50	55	...	26	29	−11	00	13	14	12	30	19	32	37	21	03	11	10	25	37
7. Mathematics A	14	25	58	43	04	26	...	57	32	47	43	11	19	21	26	16	10	04	05	05	14	42	12
8. Mathematics B	08	18	38	43	11	29	57	...	46	57	53	05	17	18	26	16	08	−02	07	04	13	49	10
9. Numerical Operations—Front	−06	04	15	06	−17	−11	32	46	...	68	50	07	25	05	20	09	01	04	16	02	12	38	−01
10. Numerical Operations—Back	−03	09	22	16	−09	00	42	57	68	...	53	08	25	10	24	13	02	03	10	04	12	41	03
11. Dial and Table Reading	09	15	28	28	00	13	43	53	50	53	...	26	44	30	37	33	20	12	17	07	18	55	22
12. Speed of Identification	19	09	11	12	08	14	11	05	07	08	26	...	47	38	22	21	14	14	15	07	06	13	18
13. Spatial Orientation I	14	14	16	13	03	12	19	17	25	25	44	47	...	46	30	27	17	14	16	06	13	33	20
14. Spatial Orientation II	22	18	22	26	16	30	21	18	05	10	30	38	46	...	25	27	20	14	11	08	08	30	26
15. SAM Discrimination Reaction Time	10	13	16	24	05	19	26	26	20	24	37	22	30	25	...	34	23	19	17	10	25	36	25
16. SAM Complex Coordination	14	11	12	22	22	32	16	16	09	13	33	21	27	27	34	...	46	32	22	16	13	27	39
17. SAM Two-Hand Coordination	13	06	06	20	34	37	10	08	01	02	20	14	17	20	23	46	...	34	20	16	12	19	35
18. SAM Rotary Pursuit	08	09	05	08	10	21	04	−02	04	03	12	14	14	14	19	32	34	...	26	21	14	04	25
19. Finger Dexterity	05	04	03	04	04	03	05	07	16	10	17	15	16	11	17	22	20	26	...	15	15	16	13
20. Aiming Stress	04	05	07	08	06	11	05	04	02	04	07	07	06	08	10	16	16	21	15	...	00	09	11
21. Bombardier Graduation-Elimination	00	04	05	13	00	10	14	13	12	12	18	06	13	08	25	13	12	14	15	00	...	17	05
22. Navigator Graduation-Elimination	11	15	33	38	08	25	42	49	38	41	55	13	33	30	36	27	19	04	16	09	17	...	16
23. Pilot Graduation-Elimination	21	12	09	20	28	37	12	10	−01	03	22	18	20	26	25	39	35	25	13	11	05	16	....

[1] Decimal points omitted.
[2] For code numbers see table 28.5.

part-whole correlations so that the communalities of these variables slightly exceeded 1.00 as should have been expected. From the comparison of loadings from this analysis with those from other analyses, it is apparent that no other serious distorting effects occurred.

The November 1943 Battery.[4]—The intercorrelations of this battery were based upon 1,900 unclassified aviation students sampled from 10 examining units. Added to the list of variables, are achievement tests in the subjects of history, geography, and physics; also two experimental tests—Decoding, CI214AX2, and Vocabulary (AAF), CI604B. The intercorrelations of these with the battery tests and with each other were based upon 543 unclassified students tested at Psychological Research Unit No. 3.

The achievement examinations had been designed to evaluate students who had completed 5 months of college training provided by the AAF, which included the three subjects mentioned along with English and mathematics. The last two subjects were covered in this battery by Reading Comprehension and Mathematics A. The objectives of the AAF English course stressed reading. Here was an opportunity to bring academic achievement into the factorial picture. This was regarded to be pertinent because of the previous discovery of the mechanical-experience factor and what seemed to be either a science-education factor or a mathematics-background factor. The latter area needed some clarification, which it was hoped would be provided by the inclusion of both physics and mathematics examinations in the matrix. It was also desired to know how much kinship existed between a physics-examination score and the mechanical factor.

This particular analysis presented one or two technical difficulties. The absence of Numerical Operations, Mechanical Information, and Speed of Identification from the battery meant the lack of three excellent reference tests. The perceptual-speed factor did emerge, but it was impossible to separate the numerical factor from general reasoning in spite of considerable effort to do so.

The presence of the vocabulary test was favorable for the appearance of the verbal factor, which rarely fails to emerge. In view of the heavy verbal loading previously found in the Technical Vocabulary and General Information tests, it was desired to know just how large the verbal factor loading would be in a nontechnical vocabulary test.

The September 1944 Battery.[5]—The intercorrelations for this battery were based upon testing of 8,158 unclassified aviation students at Medical and Psychological Examining Unit No. 8. They are presented in table 28.4. One correction was made in the coefficient of correlation between General Information, CE505F, and Mechanical Information, CI905B, in view of the fact that they had six items in common. The re-

[4] The analysis was executed by Capt. Lloyd G. Humphreys and S/Sgt. Wayne S. Zimmerman.
[5] This analysis was executed by Capt. John I. Lacey and Sgt. Harold H. Singer.

TABLE 28.2.—*Correlation matrix for the July 1943 Classification battery (N=3,000)*[1]

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16[3]	17	18	19	20	21	22	23	24
1. Reading Comprehension[2]	...	26	13	11	23	05	16	09	03	44	50	31	38	47	05	08	19	04	22	19	82	42	57	29
2. Spatial Orientation II	26	...	37	34	24	-01	04	17	05	32	20	25	15	14	05	10	22	07	23	21	41	35	37	33
3. Spatial Orientation I	13	37	...	43	39	27	24	05	09	12	17	16	21	17	11	15	28	17	24	10	26	47	40	44
4. Speed of Identification	11	34	43	...	23	11	11	14	04	13	15	26	11	04	07	12	21	16	21	14	20	45	35	33
5. Dial and Table Reading	23	24	39	23	...	44	48	-03	16	10	25	06	41	40	05	11	28	20	39	17	42	66	65	28
6. Numerical Operations, Front	05	-01	27	11	44	...	69	-23	12	-16	16	-05	35	35	00	05	05	14	16	-04	15	43	54	03
7. Numerical Operations, Back	16	04	24	11	48	69	...	-18	13	-04	21	-03	43	45	-04	02	07	11	24	01	29	48	61	05
8. Biographical Data—Pilot	09	17	05	14	-03	-23	-18	...	11	36	-02	28	-04	-08	09	14	21	09	05	32	16	15	-02	54
9. Biographical Data—Navigator	03	05	09	04	16	12	13	11	...	-06	12	-05	21	13	03	03	03	04	08	-04	09	09	32	06
10. Mechanical Principles	44	32	12	13	10	-16	-04	36	-06	...	18	27	17	26	12	20	31	08	20	41	66	47	25	53
11. General Information - Navigator	50	20	17	15	25	16	21	-02	12	18	...	31	51	36	03	04	06	02	15	00	56	29	58	14
12. General Information—Pilot	31	25	16	26	06	-05	-03	28	-05	27	31	...	04	03	05	14	17	08	07	20	33	23	17	36
13. Mathematics A	38	15	21	11	41	35	43	-04	21	17	51	04	...	56	01	03	09	04	25	04	62	42	76	15
14. Mathematics B	47	14	17	04	40	35	45	-08	13	26	36	03	56	...	-02	00	09	00	24	04	69	42	70	12
15. Aiming Stress	05	05	11	07	05	00	-04	09	03	12	03	05	01	-02	...	22	27	18	09	11	17	26	05	38
16. SAM Rotary Pursuit[2]	08	10	15	12	11	05	02	14	03	20	04	14	03	00	22	...	34	26	17	30	16	31	09	60
17. SAM Complex Coordination	19	22	28	21	28	05	07	21	03	31	06	17	09	09	27	34	...	25	35	45	33	51	24	74
18. Finger Dexterity	04	07	17	16	20	14	11	09	04	08	02	08	04	00	18	26	25	...	19	20	11	36	13	30
19. SAM Discrimination Reaction Time	22	23	24	21	39	16	24	05	08	20	15	07	25	24	09	17	35	19	...	25	45	61	39	36
20. SAM Two-Hand Coordination	19	21	10	14	17	-04	01	32	-04	41	00	20	04	04	11	30	45	20	25	...	29	53	11	61
21. Officer Quality	82	41	26	20	42	15	29	16	09	66	56	33	62	69	17	16	33	11	45	29	...	67	76	48
22. Bombardier Stanine	42	35	47	45	66	43	48	15	09	47	29	23	42	42	26	31	51	36	61	53	67	...	69	68
23. Navigator Stanine	57	37	40	35	65	54	61	-02	32	25	58	17	76	70	05	09	24	13	39	11	76	69	...	31
24. Pilot Stanine	29	33	44	33	28	03	05	54	06	53	14	36	15	12	38	60	74	30	36	61	48	68	31	...

[1] Decimal points omitted.
[2] Without divided-attention feature.
[3] For code numbers see table 28.6.

Table 28.3.—*Correlation matrix for the November 1943 classification battery (N=1,900)*[1]

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1. Biographical Data (Navigator)[2]	....	15	13	04	16	17	06	−05	11	13	09	24	11	02	−06	01	−05	09	18	13	−01	03	14
2. Biographical Data (Pilot)	15	...	05	16	−04	03	14	17	08	05	18	−02	−06	39	28	13	24	07	−21	−11	08	−16	−16
3. Spatial Orientation I	13	05	...	44	45	14	27	01	29	32	29	17	11	13	03	15	13	21	07	11	01	29	13
4. Spatial Orientation II	04	16	44	...	35	21	24	08	24	31	34	19	16	26	23	14	14	11	12	30	16	30	11
5. Dial and Table Reading	16	−04	45	35	...	36	35	04	36	52	38	42	45	13	13	18	16	23	20	28	18	51	33
6. Reading Comprehension	17	03	14	21	36	...	16	−01	21	34	32	51	50	40	31	03	05	04	44	56	57	32	51
7. SAM Complex Coordination	06	14	27	24	35	16	...	32	36	24	36	13	15	20	29	38	48	35	01	06	24	26	06
8. Rudder Control	−05	17	01	08	04	−01	32	...	07	04	19	−06	−01	17	27	32	30	11	−15	−06	18	−08	12
9. SAM Discrimination Reaction Time	11	08	29	24	36	21	36	07	...	34	31	21	23	12	20	18	23	26	12	22	23	54	16
10. Instrument Comprehension I	13	05	32	31	52	34	24	04	34	...	38	33	36	25	19	10	14	14	18	25	28	36	28
11. Instrument Comprehension II	09	18	29	34	38	32	36	19	31	38	...	24	30	36	39	24	30	13	12	22	31	38	21
12. Mathematics A	24	−02	17	19	42	51	13	−06	21	33	24	...	54	13	19	01	03	04	31	40	42	34	32
13. Mathematics B	11	−06	11	16	45	50	15	−01	23	36	30	54	...	16	27	−02	06	01	31	44	47	40	32
14. General Information	02	39	13	26	13	40	20	17	12	25	36	13	16	...	41	15	23	07	18	36	44	07	30
15. Mechanical Principles	−06	28	03	23	13	31	29	27	20	19	39	19	27	41	...	21	35	04	−02	17	50	10	00
16. SAM Rotary Pursuit	01	13	15	14	18	03	38	32	18	10	24	01	−02	15	21	...	33	33	−08	−03	05	05	04
17. SAM Two-Hand Coordination	−05	24	13	14	16	05	48	30	23	14	30	03	06	23	35	33	...	27	−12	−07	13	11	−05
18. Finger Dexterity	09	07	21	11	23	04	35	11	26	14	13	04	01	07	04	33	27	...	02	02	07	23	17
19. History	18	−21	07	12	20	44	01	−15	12	18	12	31	31	18	−02	−08	−12	02	...	72	29	22	46
20. Geography	13	−11	11	30	28	56	06	−06	22	25	22	40	44	36	17	−03	−07	02	72	...	43	29	46
21. Physics	−01	08	01	16	18	57	24	18	23	28	31	42	47	44	50	05	13	07	29	43	...	18	28
22. Decoding	03	−16	29	30	51	32	26	−08	34	36	38	34	40	07	10	05	11	23	22	29	18	...	29
23. Vocabulary (AAF)	14	−16	13	11	33	51	06	12	16	28	21	32	32	30	00	04	−05	17	46	46	28	29	....

[1] Decimal points omitted.
[2] For code numbers see table 28.7.

TABLE 28.4.—*Correlation matrix for the September 1944 classification battery (N=8,158)*[1]

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Dial and Table Reading[a]	....	03	17	40	35	41	37	20	32	52	55	02	14	27	49	14	05	22	36	22	45
2. Biographical Data (P)	03	....	20	08	13	01	19	26	13	-13	-11	35	36	13	00	13	27	08	18	21	10
3. Biographical Data (N)	17	20	....	09	04	13	08	00	08	14	12	-07	06	07	14	05	00	08	07	02	16
4. Spatial Orientation I	40	08	09	....	39	18	28	15	46	23	23	01	15	14	16	16	05	19	30	17	30
5. Spatial Orientation II	35	13	04	39	....	33	37	39	36	10	17	18	23	28	30	13	14	13	29	19	31
6. Reading Comprehension	41	01	13	18	33	....	33	32	18	29	35	18	27	41	52	04	02	10	20	09	33
7. Instrument Comprehension	37	19	08	28	37	33	....	38	28	17	22	22	36	26	30	20	23	15	36	25	36
8. Mechanical Principles	20	26	00	15	39	32	38	....	16	-02	08	47	35	33	34	21	34	07	31	33	26
9. Speed of Identification	32	13	08	46	36	18	28	16	....	15	18	07	23	14	12	15	05	20	28	14	26
10. Numerical Operations I	52	-13	14	23	10	29	17	-02	15	....	67	-13	02	13	40	03	-07	15	16	04	27
11. Numerical Operations II	55	-11	12	23	17	35	22	08	18	67	....	-06	05	20	50	04	-05	14	12	07	30
12. Mechanical Information	02	35	-07	01	18	18	22	47	07	-13	-05	....	42	25	13	08	28	04	18	28	05
13. General Information	14	36	06	15	23	27	36	35	23	02	05	42	....	29	14	16	29	08	22	19	16
14. Practical Judgment	27	13	07	14	28	41	26	33	14	13	20	25	29	....	34	08	12	08	20	14	24
15. Arithmetic Reasoning	49	00	14	16	30	52	30	34	12	40	50	13	14	34	....	01	02	09	18	11	32
16. Rotary Pursuit	14	13	05	16	13	04	20	21	15	03	04	08	16	08	01	....	35	30	39	33	23
17. Rudder Control	05	27	00	05	14	02	23	34	05	-07	-05	28	29	12	02	35	....	11	39	40	15
18. Finger Dexterity	22	08	08	19	13	10	15	07	20	15	14	04	08	08	09	30	11	....	29	23	27
19. Complex Coordination	36	18	07	30	29	20	36	31	28	16	13	18	22	20	18	39	39	29	....	49	43
20. Two-Hand Coordination	22	21	03	17	19	09	25	33	14	04	07	28	19	14	11	33	40	23	49	....	27
21. Discrimination Reaction Time	45	16	16	30	31	33	36	26	26	27	30	05	16	24	32	23	15	27	43	27	....

[1] Decimal points have been omitted.
[a] For code numbers see table 28.2.

sulting correlation between these two variables may be regarded as having that much specific overlap expunged but with other common variance unchanged.

There are two aspects of this analysis that are unsatisfactory. One is the impossibility of separating the pilot-interest and mathematical-background factors. The two Biographical Data scores come out together on the same axis, which otherwise appears to be the pilot-interest vector. Had Mathematics A been in this battery, we could have confidently expected the separation. The other discrepancy is that the Rudder Control test comes out on the psychomotor-coordination factor, which was not true in the analysis of the November 1943 battery. Although this also happened in the analysis of the carefulness battery (see ch. 25), it is believed that this test actually has a unique factor, in common with only the Rotary Pursuit test in the classification battery. What third type of test would be needed in order to effect the expected separation is not known.

The centroid factor loadings and communalities for the four analyses are given in tables 28.5, 28.6, 28.7, and 28.8. The rotated factor loadings are given in tables 28.9, 28.10, 28.11, and 28.12.

The Apparatus Tests.--In these batteries appear several apparatus tests not described previously in this volume. They are described very fully in report No. 4, so a minimum of description will be given here.

Discrimination Reaction Time, CP611D, was designed as a test of speed of decision and reaction. There are four stimulus patterns, each consisting of a pair of lights, one red and one green. Corresponding to each stimulus pattern is a micro-switch, the four switches being arranged in a diamond-shaped pattern. The position of each switch—upper, lower, right, and left—is associated with a corresponding direction of the red light with respect to the green light (see fig. 5.2 for a schematic diagram of the apparatus). The test requires 80 reactions with stimuli given in random sequence. The score is the total accumulated time between stimulus and correct response. A white signal light informs the examinee of the correctness of each response he makes.

The Finger Dexterity Test, CM116A, consists of a pegboard having 48 square pegs in square holes. Each peg can be grasped by means of a thick circular button at its top. The examinee lifts each peg from its hole, turns it 180° clockwise, and resets it in its hole. The score is the total number of pegs turned in the time allowed.

The Aiming Stress Test, CE211A, is a type of steadiness test in which the examinee tries to keep a rod delicately balanced on a fulcrum so as to avoid contact of the end of the rod with the sides of the hole into which it is inserted. During this activity, he is distracted by a "patter" which is intended to be disturbing to him, and by secondary mental tasks such as counting flashes of light. The score is the total time of contact during a given interval.

Table 28.5.—*Centroid factor loadings and communalities for the December 1942 classification battery*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	X	h^2
1. Technical Vocabulary—Pilot, CE505C	38	12	−41	−22	06	21	10	−18	08	02	47
2. Technical Vocabulary—Bombardier, CE505C	38	−07	−32	−18	16	16	05	02	16	05	37
3. Technical Vocabulary—Navigator, CE505C	53	−30	−39	26	26	−05	−12	−04	−13	−08	70
4. Reading Comprehension, CI614G	61	−13	−46	03	05	−07	15	−05	−10	32	64
5. Mechanical Information, CI905A	35	33	−48	19	−20	13	13	−15	−07	−14	61
6. Mechanical Principles, CI903A	53	25	−43	24	−19	−15	05	16	−05	−10	68
7. Mathematics A, CI702E	57	−44	−14	−05	08	−18	−20	15	−09	−09	65
8. Mathematics B, CI206B, 706A	57	−49	−05	17	−19	−06	−05	09	−08	09	66
9. Numerical Operations—Front, CI701B	38	−49	41	09	−11	26	−08	−06	08	−09	67
10. Numerical Operations—Back, CI701B	46	−53	33	11	−17	19	−15	05	08	−07	71
11. Dial and Table Reading, CP622A, 621A	65	−27	30	−02	−14	−11	−08	−13	08	07	65
12. Speed of Identification, CP610A	38	26	18	−40	−13	03	06	08	−07	−06	45
13. Spatial Orientation I, CP501B	50	11	32	−41	−19	−05	06	03	03	−08	58
14. Spatial Orientation II, CP503B	50	20	04	−32	−17	−14	05	09	−04	05	46
15. SAM Discrimination Reaction Time, CP611D	51	02	22	04	11	−17	15	03	15	09	40
16. SAM Complex Coordination, CM701A	53	32	13	16	04	−13	−15	−12	10	03	49
17. SAM Two-Hand Coordination, CM101A	44	39	04	26	05	−07	−12	−13	−07	−15	49
18. SAM Rotary Pursuit, CM803A	32	33	14	14	28	10	−13	17	03	−05	39
19. Finger Dexterity, CM116A	28	14	25	09	19	14	07	03	−09	03	24
20. Aiming Stress, CE211A	20	15	05	08	15	11	−11	09	−15	16	17
21. Bombardier Graduation-Elimination	25	−03	15	11	15	−12	24	12	14	−16	23
22. Navigator Graduation-Elimination	61	−26	14	03	−10	−15	11	−12	−10	14	55
23. Pilot Graduation-Elimination	43	35	−06	09	−07	−06	−15	−03	18	07	39

[1] Decimal points omitted.

703320—47—52

TABLE 28.6.—*Centroid factor loadings and communalities for the July 1943 classification battery*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	h^2
1. Reading Comprehension, CI614G	58	−23	−47	10	−11	16	−04	10	06	67
2. Spatial Orientation II, CP503B	45	15	[illegible]	−31	−27	03	−02	−09	−07	42
3. Spatial Orientation I, CP501B	50	10	[illegible]	−28	−15	13	−07	−18	−14	50
4. Speed of Identification, CP610A	41	21	16	−38	−20	14	10	−05	08	46
5. Dial and Table Reading, CP622-21A	61	−21	32	−03	−14	−16	06	12	−13	60
6. Numerical Operations—Front, CI701B	36	−43	50	11	11	12	21	−13	−10	67
7. Numerical Operations—Back, CI701B	44	−49	40	15	05	03	24	−14	−07	70
8. Biographical Data—Pilot, CE602D	23	41	−28	−14	20	−23	16	−14	13	47
9. Biographical Data—Navigator, CE602D	18	−16	10	−21	17	−23	−12	−06	20	25
10. Mechanical Principles, CI903A	50	24	−49	16	−17	−10	10	−12	09	65
11. General Information—Navigator, CE505D	47	−34	−26	−23	11	18	−10	23	−02	57
12. General Information—Pilot, CE505D	34	19	−29	−21	08	26	18	09	−08	40
13. Mathematics A, CI702F	56	−52	−06	−07	09	−14	−08	03	06	63
14. Mathematics B, CI206C	53	−51	−14	16	−14	−11	−04	−14	−09	65
15. Aiming Stress, CE211A	23	23	08	11	10	09	−20	−02	13	20
16. SAM Rotary Pursuit, CM803A2	36	35	07	16	23	10	−13	−08	03	37
17. SAM Complex Coordination, CM701A	55	40	09	16	06	−11	−13	11	−15	56
18. Finger Dexterity, CM116A	30	20	25	06	08	10	05	10	15	25
19. SAM Discrimination Reaction Time, CP611D	51	04	16	11	−20	−12	−05	19	03	39
20. SAM Two-Hand Coordination, CM101A	44	42	−12	21	03	−22	20	11	−09	54
21. Officer Quality	84	−23	−47	17	−22	06	−16	04	14	111
22. Bombardier Stanine	90	11	23	18	−20	−06	21	11	12	102
23. Navigator Stanine	83	−52	07	−19	−04	−09	−06	−06	06	102
24. Pilot Stanine	75	60	−04	09	25	−08	−17	−20	−10	106

[1] Decimal points omitted.

TABLE 28.7.—*Centroid factor loadings and communalities for the November 1943 classification battery*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	X	h^2
1. Reading Comprehension, CI614H	64	−39	25	−11	06	−03	04	08	−18	08	69
2. Spatial Orientation II, CP503B	48	15	−14	−16	−35	−07	19	12	11	−03	49
3. Spatial Orientation I, CP501B	42	12	−39	−12	−29	−16	08	10	−11	−14	52
4. Dial Reading—Table Reading, CP622A, 621A	63	−13	−44	−05	−07	03	−08	07	−06	−09	64
5. Biographical Data—Pilot, CE602D	17	43	19	−38	10	−28	−08	−23	10	05	55
6. Biographical Data—Navigator, CE602D	20	−12	−11	−07	08	−37	−11	−15	15	−14	29
7. Mechanical Principles, CI903B	47	26	32	−23	19	26	11	02	09	12	58
8. General Information (Pilot), CE505E	51	15	48	−19	−21	−09	−12	−12	−13	10	66
9. Mathematics A, CI702F	53	−42	−09	−19	28	−08	09	11	06	−03	61
10. Mathematics B, CI206C	55	−41	−03	−13	22	22	03	05	05	01	59
11. Instrument Comprehension I, CI615B	57	−09	−21	−14	−10	11	−18	−07	−06	−11	47
12. Instrument Comprehension II, CI616B	61	17	−04	−12	−09	15	−04	−05	11	11	48
13. SAM Rotary Pursuit, CP410B	31	43	−05	22	09	−10	−10	15	08	12	40
14. SAM Complex Coordination, CM701A	53	40	−13	21	19	06	07	−05	−06	−14	57
15. Finger Dexterity, CM116A	32	22	−19	32	08	−16	−05	−05	−21	09	38
16. SAM Discrimination Reaction Time, CP611D	50	09	−23	08	04	03	05	−15	−04	−05	35
17. SAM Two-Hand Coordination, CM101A	37	50	02	12	17	11	05	−13	−04	−39	47
18. Rudder Control, CM120B	22	38	19	14	11	13	−23	25	14	−16	44
19. History	39	−55	19	28	−21	−15	15	−15	09	−11	70
20. Geography	56	−48	27	15	−25	−09	23	−05	14	−03	79
21. Physics, CI801A	58	−13	39	−12	21	17	12	10	−09	08	63
22. Decoding, CI214AX2	51	−18	−37	07	−10	21	10	−05	−10	18	54
23. Vocabulary (AAF), CI604B	47	−38	12	23	−14	−09	−29	11	−13	−04	58

[1] Decimal points omitted.

TABLE 28.8—*Centroid factor loadings and communalities for the September 1944 classification battery*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	h^2
1. Dial and Table Reading, CP621-622A	64	−41	14	07	02	05	11	−04	08	03	[illegible]	[illegible]
2. Biographical Data (P), CE602D	30	37	−10	−28	32	−06	18	08	−05	−16	04	49
3. Biographical Data (N), CE602D	19	−12	06	−17	21	−29	12	−03	−07	−11	08	25
4. Spatial Orientation I, CP501B	47	−20	18	−32	−16	24	04	07	05	−06	08	50
5. Spatial Orientation II, CP503B	54	−03	−25	−19	−28	14	13	09	−07	−04	−03	49
6. Reading Comprehension, CI614H	53	−29	−32	11	−07	−16	08	−09	09	−04	10	54
7. Instrument Comprehension, CI616C	59	04	−10	−08	−04	06	08	−19	06	14	06	44
8. Mechanical Principles, CI903B	56	27	−35	13	−16	06	11	11	−15	05	07	61
9. Speed of Identification, CP610A	45	−12	10	−39	−13	21	−08	07	06	02	09	46
10. Numerical Operations I, CI701B	39	−53	24	24	29	12	−06	−11	−02	−06	−06	67
11. Numerical Operations II, CI701B	46	−55	13	29	26	14	−11	−06	−13	09	−05	75
12. Mechanical Information, CI905B	35	42	−39	09	09	14	−06	20	15	−04	−03	56
13. General Information, CE505F	48	28	−29	−13	15	05	−18	−16	10	11	15	54
14. Practical Judgment, CI301C	46	−07	−30	11	−09	−13	−13	02	04	−07	−01	37
15. Arithmetic Reasoning, CI206C	53	−37	−24	28	05	−11	14	02	−07	12	08	61
16. Rotary Pursuit, CP410B	37	30	34	02	−14	−09	−11	−04	−15	05	12	42
17. Rudder Control, CM120B	37	50	08	15	−02	02	10	−11	−08	−10	08	46
18. Finger Dexterity, CM116A	34	05	30	−05	−05	−14	−16	08	01	06	05	27
19. Complex Coordination, CM701A	61	22	30	06	−19	−02	10	−06	14	[illegible]	−04	60
20. Two-Hand Coordination, CM101A	48	37	22	19	−08	02	11	11	12	−04	−04	50
21. Discrimination Reaction Time, CP611D	59	−12	16	−03	−10	−19	09	−06	06	12	−11	48

[1] Decimal points omitted.

TABLE 28.9.—*Rotated factor loadings for the December 1942 classification battery*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	X	h^2
1. Technical Vocabulary—Pilot[2]	41	17	−08	−01	01	39	−01	05	34	−00	47
2. Technical Vocabulary—Bombardier	44	08	04	−05	15	15	00	12	33	01	37
3. Technical Vocabulary—Navigator	77	11	14	08	15	10	06	01	00	05	67
4. Reading Comprehension	54	03	09	16	25	40	−03	10	06	27	63
5. Mechanical Information	10	00	−09	05	15	74	04	−01	09	04	60
6. Mechanical Principles	14	04	−02	19	52	56	09	−02	−00	15	67
7. Mathematics A	53	07	42	16	33	−04	06	01	−12	12	63
8. Mathematics B	28	−01	59	14	25	09	−03	02	−02	36	64
9. Numerical Operations—Front	04	08	76	02	−10	−09	08	22	06	05	66
10. Numerical Operations—Back	09	07	81	03	06	−07	07	14	05	09	70
11. Dial and Table Reading	14	30	56	40	07	−02	05	14	11	19	65
12. Speed of Identification	02	63	00	03	09	09	12	11	07	02	45
13. Spatial Orientation I	02	69	20	16	10	01	04	16	07	02	58
14. Spatial Orientation II	11	56	02	17	24	12	04	03	09	14	46
15. SAM Discrimination Reaction Time	07	19	14	36	20	−03	10	35	16	18	42
16. SAM Complex Coordination	−01	17	09	49	17	20	34	10	17	02	50
17. SAM Two-Hand Coordination	−03	08	02	39	12	37	40	10	−02	−04	48
18. SAM Rotary Pursuit	−02	07	−04	13	20	06	32	20	14	−06	40
19. Finger Dexterity	−03	11	−06	07	−03	03	36	28	07	12	25
20. Aiming Stress	04	03	−03	03	05	02	36	00	10	17	18
21. Bombardier Graduation-Elimination	00	03	08	14	20	−01	02	43	−03	00	25
22. Navigator Graduation-Elimination	21	21	39	34	06	09	−01	17	01	38	54
23. Pilot Graduation-Elimination	−02	15	01	34	25	26	22	−03	28	−02	39

[1] Decimal points omitted.
[2] For code numbers see table 28.5.

Table 28.10.—*Rotated factor loadings for the July 1943 classification battery*[1].

Test	I	II	III	IV	VI	VII	IX	X	XI	h^2
1. Reading Comprehension[2]	66	−01	15	16	34	−01	01	26	05	67
2. Spatial Orientation II	20	46	−05	12	18	−06	25	23	08	43
3. Spatial Orientation I	20	48	19	04	00	10	38	19	05	50
4. Speed of Identification	11	63	07	03	11	02	18	00	01	46
5. Dial and Table Reading	10	28	51	43	−03	−01	16	13	15	60
6. Numerical Operations, Front	03	13	79	−05	−09	09	−10	02	−06	67
7. Numerical Operations, Back	02	05	83	66	01	−01	07	12	−05	72
8. Biographical Data—Pilot	−05	13	−19	−05	50	22	11	−09	29	46
9. Biographical Data—Navigator	−01	12	17	−09	−04	04	−12	07	42	25
10. Mechanical Principles	26	09	−12	26	64	10	04	27	04	65
11. General Information—Navigator	63	05	22	−02	02	−07	13	06	27	55
12. General Information—Pilot	40	18	−06	−04	30	08	32	−17	00	43
13. Mathematics A	37	03	51	07	07	−07	−01	25	37	61
14. Mathematics B	27	−07	48	19	16	−14	02	48	12	63
15. Aiming Stress	07	12	−03	07	00	39	−05	06	00	18
16. SAM Rotary Pursuit	06	10	02	08	12	56	11	05	01	36
17. SAM Complex Coordination	07	18	01	47	11	46	22	06	12	55
18. Finger Dexterity	07	26	15	16	03	33	−05	−14	−04	25
19. SAM Discrimination Reaction Time	15	26	22	48	06	14	−02	09	08	41
20. SAM Two-Hand Coordination	−01	08	−01	45	44	27	18	−09	09	52
21. Officer Qualification	71	10	23	33	42	05	−07	48	13	111
22. Bombardier Stanine	19	47	48	58	35	26	03	06	03	102
23. Navigator Stanine	41	31	67	15	09	−07	04	37	35	101
24. Pilot Stanine	06	29	−05	27	37	73	38	20	23	107

[1] Decimal points omitted.
[2] For code numbers see table 28.6.

TABLE 28.11.—*Rotated factor loadings for the November 1943 classification battery*[1]

Test	I	II	IV	VI	VII	XI	XII	XIII	—	Res	h^2
1. Reading Comprehension[2]	58	09	−02	33	−02	05	04	10	48	09	70
2. Spatial Orientation II	02	56	24	11	−08	11	18	20	13	12	51
3. Spatial Orientation I	03	62	17	−09	03	23	07	00	18	00	51
4. Dial Reading-Table Reading	23	38	36	−14	06	25	02	−02	46	09	64
5. Biographical Data—Pilot	−12	10	22	53	−03	31	−08	−15	−26	03	55
6. Biographical Data—Navigator	18	02	05	−01	06	30	−08	05	−03	−01	30
7. Mechanical Principles	03	01	38	58	03	−12	17	03	20	06	57
8. General Information (Pilot)	43	29	23	53	−08	−08	−04	−06	−14	07	65
9. Mathematics A	30	−02	08	18	02	36	05	16	56	07	61
10. Mathematics B	36	−08	26	12	−04	12	03	14	58	09	60
11. Instrument Comprehension I	29	27	40	22	−05	15	−02	−10	30	05	45
12. Instrument Comprehension II	14	28	50	12	01	03	09	02	16	22	48
13. SAM Rotary Pursuit	−05	16	23	−03	42	03	27	−07	−10	21	40
14. SAM Complex Coordination	−01	22	46	15	47	04	14	01	11	−17	57
15. Finger Dexterity	10	22	15	−05	51	03	−05	−08	02	02	35
16. SAM Discrimination Reaction Time	10	24	39	02	24	10	−05	07	19	−03	33
17. SAM Two-Hand Coordination	00	15	44	30	36	−06	10	−01	−03	−16	47
18. Rudder Control	09	−03	28	01	10	01	51	−16	23	−05	44
19. History	63	08	−02	−08	−02	08	−10	52	08	−05	70
20. Geography	61	15	06	12	−08	01	01	58	16	06	79
21. Physics	38	−03	17	51	04	−11	11	11	38	00	61
22. Decoding	14	31	32	−17	12	−06	−15	10	46	20	55
23. Vocabulary (AAF)	71	11	04	−11	10	04	09	−01	16	11	59

[1] Decimal points omitted.
[2] For code numbers see table 28.7.

TABLE 28.12.—*Rotated factor loadings for the September 1944 classification battery (N=8,158)*[1]

Test	I	II	III	IV	V	VI	VII	VIII	IX	X	Residual	h^2
1. Dial and Table Reading, CP621–632A	12	30	33	40	00	−02	10	05	06	24	−01	62
2. Biographical Data (P), CE602D	−07	07	−10	01	18	38	04	16	49	00	12	48
3. Biographical Data (N), CE602D	00	02	14	04	−03	−10	−03	20	37	13	12	24
4. Spatial Orientation I, CP501B	02	01	21	16	05	00	14	04	08	04	−06	48
5. Spatial Orientation II, CP503B	16	49	07	21	33	10	01	03	00	20	04	48
6. Reading Comprehension, CI614H	52	08	31	22	15	00	−02	−02	09	28	−02	53
7. Instrument Comprehension, CI616C	17	20	11	41	24	14	13	02	17	13	−17	42
8. Mechanical Principles, CI903B	20	09	−02	17	51	37	15	00	−03	31	02	59
9. Speed of Identification, CP610A	08	58	14	09	08	05	11	10	11	00	−18	45
10. Numerical Operations I, CI701B	00	05	76	25	−08	−10	00	00	02	−02	04	66
11. Numerical Operations II, CI701B	03	02	81	23	06	07	−03	03	−07	08	−08	74
12. Mechanical Information, CI905B	26	00	−08	02	23	64	03	−02	02	01	01	54
13. General Information, CE505F	37	06	03	10	26	34	16	−03	36	−09	−15	52
14. Practical Judgment, CI301C	46	03	17	13	22	12	00	07	01	16	05	36
15. Arithmetic Reasoning, CI206C	24	−02	46	27	18	.04	−07	00	00	47	00	60
16. Rotary Pursuit, CP410B	−03	05	−01	14	18	00	53	27	00	−02	−01	41
17. Rudder Control, CM120B	−03	−05	−11	22	28	27	43	01	11	−02	14	44
18. Finger Dexterity, CM116A	06	13	12	10	−02	−01	31	34	00	00	−03	26
19. Complex Coordination, CM701A	09	25	05	47	08	16	44	17	01	00	18	58
20. Two-Hand Coordination, CM101A	00	08	00	32	08	36	41	16	−07	03	19	48
21. Discrimination Reaction Time, CP611D	14	20	20	47	07	53	13	31	04	17	01	47

[1] Decimal points omitted.

The Rotary Pursuit Test, CM803A, is a modified Koerth pursuit test in which the examinee tries to keep a prod in contact with a metallic spot on a phonograph-type disk which is rotating at the rate of one revolution per second. In the same test with divided attention, CP401B, there is a second simultaneous task for the left hand which requires the examinee to keep one of two keys closed in correspondence with one of two lights.

In the Rudder Control Test, CM120B, the examinee sits in a mock cockpit of an airplane. His own weight throws the seat off balance unless he applies correction by means of a rudder-control mechanism. The score is the total time he keeps the cockpit pointed directly at a target light straight ahead. The task requires a keen appreciation of loss of balance and a quick but not over-controlled correction made by leg action.

The Factors

The statistical results from these analyses are summarized factor by factor in the following paragraphs. The tests as well as the factors are very much in common to all four analyses. It should be noted, however, that several tests changed form with change of battery. In December 1942, the General Information Test was the Technical Vocabulary and Information Test, CE505C; in July 1943, it was General Information, CE505D; in November 1943, it was General Information, CE505E; and in September 1944, it was CE505F. Reading Comprehension was form CI614G in the first two batteries and CI614H in the last two. Mathematics A was CI702E in the first battery and CI702F in the second and third. Mathematics B provided an unweighted combination of two scores, one from CI206B (Arithmetic Reasoning) and one from CI706A (Numerical Approximation), in the first battery. The score in the last three batteries was derived from a single test, CI206C (Arithmetic Reasoning). It was Mechanical Principles, CI903A, in the first two batteries, and CI903B in the last two. It was Rotary Pursuit, CM803A (without the divided attention feature) in the first battery and CP410B (with divided attention) in the last three batteries.

In the tabulations below, no test is listed in any group unless its loading for the factor exceeds 0.20 in all batteries in which it had a loading at all. A blank means that a test or a score was absent from a particular analysis. For the November 1943 analysis, no data are given for the numerical or general-reasoning factors because of the failure to separate the two.

Factor I is the common verbal factor. All repeated estimates for any test are rather consistent with one or two exceptions. The drop from a loading of 0.53 in Mathematics A to 0.37 and 0.29 coincides with a change of form, which might indicate that the new form lost much of

Rotated factor I is described by the following data:

Test name	Factor loadings			
	December 1942 battery	July 1943 battery	November 1943 battery	September 1944 battery
General Information (navigator score)	0.77	0.63	....	....
Reading Comprehension	.54	.66	0.57	0.52
Mathematics A	.53	.37	.29	....
Technical Vocabulary and Information (bombardier score)	.44	....	....	....
General Information (pilot score)	.41	.40	.43	.37
Mathematics B	.28	.27	.35	.24
Instrument Comprehension I	....	....	.29	....
Vocabulary, AAF	....	....	.71	....
Practical Judgment	....	....	....	.46

its verbal variance. It should also be noted, however, that another factor —mathematical background—did not emerge in the December 1942 analysis. It is possible that its variance became embroiled with the verbal variance in that analysis. The drop from 0.77 to 0.63 in going from Technical Vocabulary and Information, CE505C, to General Information, CE505D, navigator score, is difficult to explain, since the change was in name of the test only, so far as the navigator score was concerned.

It is interesting to note that a technical vocabulary test can have as high a verbal loading as a nontechnical vocabulary test. The selection of items valid for the navigator, however, probably accounts for this fact, since the verbal loading for the pilot score is only approximately 0.40. The navigator criterion has a positive verbal loading, whereas the pilot criterion has not (see table 28.14). The selection of items correlated with the two criteria would therefore yield different results with respect to total-score verbal variance.

Rotated factor II is described by the following data:

Test name	Factor loadings			
	December 1942 battery	July 1943 battery	November 1943 battery	September 1944 battery
Spatial Orientation I	0.69	0.48	0.62	0.61
Speed of Identification	.63	.63	....	.58
Spatial Orientation II	.56	.46	.56	.49
Dial and Table Reading	.30	.28	.38	.30
Instrument Comprehension II	....	....	.28	.30
Instrument Comprehension I	....	....	.27	....

This is the familiar perceptual-speed factor with the usual loadings in the same tests.

Rotated factor III is described by the following data:

Test name	Factor loadings		
	December 1942 battery	July 1943 battery	September 1944 battery
Numerical Operations (B)	0.81	0.83	0.81
Numerical Operations (F)	.76	.79	.76
Mathematics B	.59	.48	.46
Dial and Table Reading	.56	.41	.53
Mathematics A	.43	.31	....

This is the common numerical factor with its usual very consistent loadings. Two important results should be pointed out here. One is that the back of the Numerical Operations test sheet has consistently higher loadings than the front in all analyses. The back of the sheet is composed of problems in subtraction and division, whereas the front is composed of problems in addition and multiplication. The back also provides five alternative responses, whereas the front provides only two alternatives. Which of these distinctions is responsible for the finding is not clear. The other notable feature is the drop in loading for Mathematics B coincident with the dropping of Numerical Approximations from that test. Numerical Approximations is much more of a computations task than is Arithmetic Reasoning.

It should also be remarked that when front and back scores are combined and analyzed as a single test variable, the numerical loading drops to the region of 0.65 to 0.70. This must mean that there is a specific variance common to the two parts of the Numerical Operations test in addition to their common variance in the numerical factor. In the analyses reported here the two variances have probably combined.

Rotated factor IV is described by the following data:

Test name	Factor loadings			
	December 1942 battery	July 1943 battery	November 1943 battery	September 1944 battery
Complex Coordination	0.49	0.47	0.46	0.47
Dial and Table Reading	.40	.43	.36	.40
Two-Hand Coordination	.39	.45	.44	.32
Discrimination Reaction Time	.36	.48	.39	.47
Instrument Comprehension II	....	....	.50	.41
Instrument Comprehension I	....	....	.40	....
Rudder Control	....	....	.28	.22

This is the spatial-relations (space I) factor of which the Complex Coordination test has usually been a leading measure. There is some indication that printed tests, such as Instrument Comprehension II, will provide an equally good measure of the factor. The surprising strength of correlations between the psychomotor tests in this group with certain printed tests is mainly accounted for by this common factor.

Rotated factor V is described by the following data:

Test	Factor loadings	
	December 1942 battery	September 1944 battery
Mechanical Principles	0.52	0.51
Mathematics A	.33	....
Spatial Orientation II	.34	.33

This is the visualization factor usually prominent in tasks in which some pictorial content must be mentally manipulated or transformed. Since its only strong loading in these batteries is in Mechanical Principles, it does not appear in all four analyses. It has appeared in a number of experimental battery analyses, however, where other strong tests in it have been present.

Rotated factor VI is described by the following data:

Test name	Factor loadings			
	December 1942 battery	July 1943 battery	November 1943 battery	September 1944 battery
Mechanical Information	0.74	....	....	0.64
Mechanical Principles	.56	0.64	0.58	.37
Reading Comprehension	.40	.34	.33	.00
General Information (pilot score)	.39	.30	.53	.34
Two-Hand Coordination	.37	[illegible]	.30	.36
Biographical Data (pilot score)	....	.50	.53	.33

This is the mechanical-experience factor which is exceptionally strong in the Mechanical Information test. Estimates of its loading in this test are even higher in other analyses. The Mechanical Principles test outstrips it in validity for pilot training, because it includes other valid factors also, such as visualization and spatial relations. One noteworthy fact is that four tests that were not designed as mechanical tests turn out to have substantial loadings in this factor. One reason that probably applies to two of these tests is that these tests were developed by selection of items that correlate with the pilot criterion. Item validation thus tends to work toward complicating a test factorially rather than purifying it unless one also selects or rejects items that have been validated against pure factor criteria. Thus, in selecting new items for Biographical Data, one might make sure that their correlation with the score on Mechanical Information is very low.

Rotated factor VII is described by the following data:

Test name	Factor loadings			
	December 1942 battery	July 1943 battery	November 1943 battery	September 1944 battery
Rotary Pursuit	0.52	0.56	0.42	0.53
Two-Hand Coordination	.40	.27	.36	.41
Aiming Stress	.36	.39	....	....
Finger Dexterity	.34	.33	.51	.31
Complex Coordination	.34	.46	.47	.44

This is a factor confined to psychomotor tests and so its interpretation must be given accordingly. The best name at present seems to be that of psychomotor coordination. It is strongest in the Rotary Pursuit test, but also substantial in Complex Coordination and others. It is absent from the Rudder Control test, which requires coordination of leg-muscle action, in the November 1943 analysis but not in the September 1944 analysis (see table 28.12).[a] If the former result is confirmed, the indication is that this factor is characteristic of arm-muscle activity but not of leg-muscle activity. If the latter result is confirmed, it is a more

[a] In the analysis of the Carefulness Battery (see ch. 25) the Rudder Control test also had a substantial loading with this factor.

general motor factor. Since it is substantial in one-armed activity, as in Finger Dexterity, Rotary Pursuit, and Aiming Stress, it is not a matter of coordinating the two arms. General muscular agility seems to be the best characterization of it.

Rotated factor VIII is described by data from two analyses:

Test name	Factor loadings	
	December 1942 battery	September 1944 battery
Discrimination Reaction Time	0.35	0.31
Finger Dexterity	.28	.34

This factor would probably have failed to emerge in the December 1942 analysis had not the bombardier criterion been included in the matrix. The bombardier criterion has a loading of 0.43 in the factor. It had been found in previous estimations that a number of tests, including the two listed above, and Rotary Pursuit, had bombardier validities in excess of the amount that could be attributed to other known factors. This factor sufficiently accounts for the remaining validities of these tests. It has been called the bombardier factor, but to define it psychologically, it seems to be a psychomotor-precision ability of some kind. Further work toward the improvement of selection of bombardiers should stress this factor very heavily.

Rotated factor IX is described by the following data from two tests:

Test name	Factor loadings		
	December 1942 battery	July 1943 battery	September 1944 battery
General Information (pilot score)	0.34	0.32	0.36
General Information (bombardier score)	.33	....	....

There is a very slender basis for the interpretation of this factor, but since the General Information tests (pilot score) were designed particularly to measure pilot interest and since the pilot criterion has a loading in the factor, some credence may well be given to the hypothesis that this is a pilot-interest factor. The fact that this factor has small loadings for Complex Coordination, and in one analysis for Spatial Orientation I and II, three tests which have considerable face validity for pilot, lends support to this hypothesis. If the hypothesis is correct, it would seem that the bombardier score of Technical Vocabulary and Information missed its aim. Pilot interest and bombardier interest may be closely akin, or there may be no well formed entity that can be called bombardier interest. There were suggestions that this factor be called aviation interest, but the navigator criterion seems to have little or no communality with it (see table 28.14).

Rotated factor X is described by the following data involving two tests only:

Test name	Factor loadings		
	December 1942 battery	July 1943 battery	September 1944 battery
Mathematics B	0.36	0.48	0.47
Reading Comprehension	.27	.26	.28

This combination of tests and these loadings coupled with other experience force us to accept it as the general-reasoning factor (R_1). It comes out repeatedly in a number of experimental tests not represented in these analyses. The lower loading of R_1 for Mathematics B in the December 1942 analysis is probably due to the presence of the Numerical Approximation test, which seems much more numerical and less of a reasoning test, at least by inspection.

Rotated factor XI is described by the following data:

Test name	Factor loadings	
	July 1943 battery	November 1943 battery
Biographical Data (navigator score)	0.42	0.50
Mathematics A	.37	.36
Biographical Data (pilot score)	.39	.31
General Information (navigator score)	.27	....

This factor might be called a navigation-interest factor, but it is doubtful whether such interest has sufficiently crystallized in the young men who took the examinations as to represent an entity. In view of the loading in Mathematics A, it is more likely to represent a mathematical-background factor. One hypothesis considered was that it represents a natural-science education factor, but the absence of the physics test in this list fairly well disproves that hypothesis. The factor will therefore be called mathematical background.

Rotated factor XII is described by two tests in only one analysis:

Test name	*Factor loadings Nov. 1943 battery*
Rudder Control	0.51
Rotary Pursuit	0.27

This factor is almost entirely confined to the Rudder Control test which is relatively pure with respect to it. By inspection, the Rudder Control test seems to involve motor coordination controlled mostly by the kinesthetic sense. The best hypothesis, therefore, seems to be that it is a kinesthetic-motor factor. There is no doubt of the unique contribution of this factor to pilot validity.

It should be added that this factor did not come out in other analyses (September 1944 battery and Carefulness battery). Instead, the Rudder Control test then acquired substantial loading in the psychomotor-coordination factor. In still another attempt at analysis of the November 1943 battery combined with additional tests, the kinesthetic factor again appeared, and the test had a zero loading for the coordination factor.

The existence of the kinesthetic factor, therefore, rests upon an insecure basis as yet. That the test in question contains much valid variance not accounted for by its correlations with other tests cannot be questioned. That some or all of this is a kinesthetic factor still needs confirmation.

Rotated factor XIII is described by two tests in one analysis:

Test name	*Factor loadings July 1943 battery*
Geography	0.58
History	.52

For all that we know from these results, this could be a doublet confined to these two tests. It is probably safe to assume that it is a more general factor, however, and the hypothesis is offered that it is a social-science background factor. This places it in a class with the mathematical-background factor and the mechanical-experience factor, which seem to represent variables in individual differences produced by learning. This would not preclude, however, the operation in each of them of inborn inclinations.

Factor Loadings of Composites

Of special interest in the July 1943 battery are the factor loadings in the composite scores or stanines. These could be estimated from the loadings of factors in the tests that enter into them and the weights assigned to the tests. The loadings, presented separately in table 28.13, are undoubtedly inflated, as was said before, due to the correlation of error variances, but their relative positions are probably correct. From table 28.13 it will be seen that the leading factor in the bombardier composite is spatial relations, other relatively strong ones being perceptual speed, numerical, and mechanical experience. The leading factor in the navigator composite is numerical, with substantial loadings in verbal, perceptual speed, general reasoning, and mathematical background. For the pilot composite, psychomotor coordination leads by a substantial margin; but the composite also has strong loadings in pilot interest, mechanical experience, and preceptual speed. The officer-quality composite is predominantly verbal, with moderate loadings in general reasoning and mechanical experience, as had been intended.

TABLE 28.13.—*Factor loadings of four composite scores, derived from the July 1943 battery analysis*

Factor	Bombardier composite	Navigator composite	Pilot composite	Officer quality
Verbal	0.19	0.41	0.06	0.71
Perceptual	.47	.31	.39	.10
Numerical	.48	.67	−.05	.23
Spatial relations	.58	.15	.77	.33
Mechanical experience	.35	.09	.37	.43
Psychomotor coordination	.36	−.07	.73	.05
Pilot interest	.03	.04	.38	−.07
General reasoning	.06	.37	.30	.48
Mathematical background	.03	.35	.23	.13

Validities of the Factors

The validities of the factors for bombardier, navigator, and pilot training are the factor loadings in the three training criteria. The loadings for the most commonly recurring factors have been estimated in three ways, one of which depended upon a least-square type of solution, one upon an iterative method, and one upon a direct analysis of the December 1942 battery.[7] The first two procedures employed data based primarily upon the July 1943 battery. All three arrived at very similar validities. In the iteration procedure, no factor loadings were permitted to be so large that test validities were significantly overestimated. Validities were permitted to be fully estimated for tests in which the factors considered also practically account for the nonchance variance of the test. For other tests the validity was permitted to be grossly underestimated, if necessary. This was to allow for the existence of unknown valid factors. In the least-square solution, as much overprediction was permitted as underprediction. The results of the three estimation methods are given in table 28.14.

TABLE 28.14.—*Estimated factor validities for bombardier, navigator, and pilot training criteria*

Factor	Least-square solution			Iterative solution			Factor-analysis solution		
	B	N	P	B	N	P	B	N	P
Verbal	0.09	0.18	0.04	0.09	0.18	0.03	0.00	0.21	−0.02
Perceptual speed	.11	.27	.14	.11	.22	.15	.03	.21	.15
Numerical	.15	.42	.00	.13	.44	.00	.08	.39	.01
Spatial relations	.22	.33	.26	.25	.33	.21	.14	.34	.34
Visualization	.....	.....	.....	.00	.06	.12	.20	.06	.25
Mechanical experience	.01	.10	.32	.01	.08	.32	−.01	.09	.26
Psychomotor coordination	.10	−.04	.19	.01	−.05	.12	.02	−.01	.22
Psychomotor precision	.....	.....	.....	.....	.....	.....	.43	.17	−.03
Pilot interest	−.11	.01	.25	−.11	.00	.20	−.03	.01	.28
General reasoning	.02	.28	−.01	.02	.25	−.01	.00	.38	−.02
Mathematical background	−.02	.26	.15	−.02	.30	.10	.....	.....	.....

The agreements among the estimates in table 28.14 are generally good, considering the facts that three factors were not in common throughout all solutions, that somewhat different principles and procedures underlie the estimates, and that somewhat different data were used as bases. The major discrepancies are worthy of comment. The small verbal and perceptual loadings for bombardier found by the first two methods virtually dropped out in the third, and those for numerical and spatial relations became greatly reduced. The introduction of the psychomotor precision factor in the December 1942 analysis seems to have been at the expense of spatial relations loadings in some tests as well as in the bombardier criterion, but did not lower variances systematically for navigator or bombardier in the verbal, perceptual, numerical, and spatial relations factors. Visualization seems to have been grossly underestimated

[7] For a discussion of these procedures in principle see report No. 3 of this series. The validity figures used in all three procedures were the same as those described in connection with the analyses of the December 1941 battery (see page 799).

for the bombardier, psychomotor coordination for the pilot, and general reasoning for the navigator, in the iterative method in particular. Further space will be given to the factor composition of the pilot criterion later in this chapter.

A comparison of tables 28.13 and 28.14 is of some interest. The weightings of tests in the composites had been derived empirically, closely in accordance with the principles of the multiple-regression equation. It might be expected, therefore, that the relative weights effective for the factors would correspond roughly with the factor loadings in the criteria. This proves to be true, with one or two notable exceptions for the pilot. Psychomotor coordination holds fourth rank for factor loading in the pilot criterion, but is given highest weight in the pilot composite. This would indicate an overweighting of psychomotor tests for the pilot in the July 1943 stanine. The spatial-relations factor has either first or second rank for weight in the pilot criterion, but holds only fifth place in the composite. Better estimates of the factor loadings of the pilot criterion (see table 28.17) do not alter the situation just described. These discrepancies could be corrected by giving psychomotor tests, such as Rotary Pursuit, less weight, and giving Dial and Table Reading more weight. An even better solution would be to purify the tests. Too many tests like Complex Coordination, Discrimination Reaction Time, and Two-Hand Coordination have substantial loadings in both the psychomotor-coordination and spatial-relations factors so that to increase the weight of one is to increase the weight of the other. In a battery of pure tests, there is much more freedom of action in arriving at optimal weights.

A SUMMARY OF FACTORIAL RESULTS

Two Master Tables

For the convenience of the reader, factorial results have been summarized so as to present the picture of tests in clearest form.

One summary is a reference list, with tests given in alphabetical order (see appendix B) and the other presents tests grouped by factors (table 28.15). The latter includes all tests analyzed as reported in earlier chapters. The former includes only the printed tests that are the primary subject matter of this volume. Wherever any test has been analyzed more than once, a weighted mean of its loadings in each factor has been determined. Three analyses have been om[itt]ed from consideration in this, the November 1943 and September 1944 battery analyses, and the carefulness battery analysis. In two instances the failure to separate factor R_1 from other factors left some uncertainty as to other aspects of those analyses. The September 1944 analysis was completed too late to be included. Each observed loading was weighted by the number of cases in the sample analyzed. It is recognized both that there is no precedent

703330—47—53

for the procedure, and that inaccuracies in rotations may far outweigh sampling errors in some instances. On the other hand, repeated estimates of the same factor loading in the same test appear to behave like sampling statistics, and to this extent the procedure seems justified. Wherever vacancies would otherwise have occurred in either summary table, single estimates of factor loadings from the November 1943 and carefulness analyses were used if they did not come from the doubtful columns of loadings.

In both summary tables, one set of loadings, one communality, one estimate of reliability, and one of validity for each criterion have been listed. The communality is the sum of the squares of loadings as listed. This value may be higher than that found in any single analysis because all factors in the test did not necessarily appear in all analyses. There is also the possibility that factor loadings from different analyses are not properly identified as pertaining to one and the same factor, though it is believed that this possibility is rather remote. The reliabilities have been determined by various procedures: Kuder-Richardson, equivalent halves, odd-even, or part I-part II intercorrelations without intervening time intervals. It is believed that the best type to compare with communalities is the fourth, which is essentially an alternate-forms procedure without intervening time interval. This method was most commonly employed in the program.

If for any test h^2 is equal to r_{11}, it means that the entire nonerror variance is accounted for by known common factors. If there is a substantial positive discrepancy ($r_{11}-h^2 > 0.00$), it means that some common factors have not yet been brought to light in the test. A substantial negative discrepancy ($r_{11}-h^2 < 0.00$), would mean that there is an error of estimation, but whether in the derivation of r_{11} or of h^2 is unknown. In both summaries the sizes of samples, upon which factorial results on the one hand and validity coefficients on the other are based, are given in two separate columns under the symbols N_f and N_v, respectively. Validities for pilot (P), bombardier (B), and navigator (N) only are given.

The more meaningful summary (table 28.15) is arranged by factors. Each test is placed in a list according to the factor in which it has its highest loading. In each list, the tests are arranged in descending order of loading in the factor. In this manner one can decide at a glance which tests are purest and strongest in each factor and which ones have similar secondary loadings. In a few instances, tests have been placed in more than one list because they have similar high loadings in two factors. The list in appendix B has a distinct advantage for those who wish to look up a specific test by name. The list in this chapter is functionally superior for those who wish to select batteries for different purposes or who wish to select alternative tests by equivalent factorial configuration.

Key to the Factors

As a supplement to these two master lists, it is desirable to present here, together in one place, definitions of factors and the symbols that are used to designate them. This list is alphabetical for the sake of easy reference.

Ca—The carefulness factor occurred in tests designed as carefulness tests, but curiously enough, and after all reasonably enough, it was strong in the error scores rather than in the rights scores.

I_1—The first integration factor is common to tests that require the effective memory of a number of rules in the carrying out of simple tasks on paper.

I_2—This is the second integration factor which is common to Following Directions tests and others in which mental sets change frequently.

I_3—The third integration factor seems to be common to tests in which the grasp of a wide variety of details is important.

J—The judgment factor is found in tests of practical judgment and practical estimations. The ability probably involves making wise choices from a number of alternative solutions to a practical problem. It seems to be a judicial or criticizing function.

K—A kinesthetic factor of some kind, as yet it is fairly specific to the rudder control test. It is not listed in table 28.15.

LE—This is a length-estimation factor involving the comparison of lines or simple distances between points. It may involve more complex estimates than those of linear dimensions.

M_1 (PM)—This memory factor has been identified as paired-associates memory. It is involved in tasks requiring the memorization of items in pairs and is evaluated by an immediate test of retention and recognition.

M_2 (VM)—The second memory factor is identified as a visual-memory ability. It is prominent in tests requiring the retention and recall of a pictorial stimulus after very short time intervals. The length of time interval may be an irrelevant condition in this as in factor M_1.

M_3—This memory factor seems to be restricted to memorizing paired-associates material in which one item is a pictorial symbol and the other is a verbal symbol.

MB—Mathematical background, which may include mathematical interest as well as mathematical training.

ME—The mechanical-experience factor is most heavily and purely weighted in the Mechanical Information Test, and tests of Driving Skill and of Tool Function.

P—The perceptual-speed factor involves the rapid comparison of visual forms, and the notation of similarities and differences in form and detail.

TABLE 28.15—*Factor loadings, communalities, reliabilities, and*

Test and Code No.	N_t	Co	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
CAREFULNESS															
Plotting Test, CE452A. (Total Wrongs)	354	59	....	....	....	....	....	....	....	....	....	....	00	....	....
Complex Scale Reading, CE454A........ (Total Wrongs)	354	57	....	....	....	....	....	....	....	....	....	....	09	....	....
Plotting Accuracy, CE453A.......... (Total Wrongs)	354	51	....	....	....	....	....	....	....	....	....	....	22	....	....
Directional Plotting, CE455A.......... (Total Wrongs)	354	41	....	....	....	....	....	....	....	....	....	....	08	....	....
INTEGRATION I															
Signal Interpretation, CI656AX2.........	266	....	59	06	30	....	10	....	-02	....	....	....	05	06	....
Combat Planes, CI655AX5.........	266	....	57	12	28	....	10	....	-03	....	....	....	22	02	....
Flight Formations, CI654AX5.........	266	....	46	14	21	....	13	....	-02	....	....	....	09	11	....
INTEGRATION II															
Following Directions, CP402A...........	266	....	11	54	10	....	05	....	-03	....	....	....	25	09	....
Code Analysis, CI653AX2.........	260	....	03	40	348	....	00	....	-07	....	....	....	29	06	....
INTEGRATION III															
Planning Air Maneuvers, CI408AX3....	638	....	01	06	43	-11	-19	....	20	....	....	....	08	06	....
Code Analysis, CI653AX2.........	260	....	03	340	42	....	00	....	-07	....	....	....	29	06	....
Planning a Course, CI406AX2.........	436	....	06	17	41	10	16	....	00	....	....	....	30	05	....
Figure Classification, CI213AX1.........	202	....	....	....	38	....	....	....	....	....	....	....	05	01	....
Spatial Reasoning, CI211BX1.........	404	....	....	....	38	14	....	....	02	....	....	....	18	16	....
JUDGMENT															
Practical Judgment II, CI301BX3......... (Work Plan)	170	....	....	....	01	45	....	....	32	....	....	....	15	-05	....
Practical Judgment, CI301BX1......... (Non-Mechanical)	202	....	....	....	....	39	....	....	13	....	....	....	15	16	....
Sequence of Maneuvers, CI410A............	202	....	....	....	....	38	....	....	00	....	....	....	30	00	....
Judgment (Pure) Commonsense, AAFQE JR P-3..........	1713	....	....	....	....	37	....	....	14	....	....	....	....	....	....
Practical Estimations I, CI308AX1.......	170	....	....	....	01	36	....	....	33	....	....	....	13	-05	....
Competitive Planning, CI409AX2.........	372	....	....	....	33	36	....	....	-02	....	....	....	15	05	....
LENGTH ESTIMATION															
Pattern Assembly, CP804A...........	202	....	....	....	....	....	52	....	02	....	....	....	....	31	....
Shorter Path—Path Distance, CP608B...	545	....	....	....	....	....	46	....	-01	....	-11	....	17	25	....
Shorter Line—Line Lengths, CP608B...	545	....	....	....	....	....	44	....	04	....	11	....	09	10	....
Nearest Point—Point Distance, CP607B...	545	....	....	....	....	....	43	....	11	....	23	....	-04	09	....
Map Distance, CP626B	858	....	01	....	-09	....	30	....	17	....	04	....	09	01	....

. . .

validities for tests grouped according to highest factor loadings[1]

FACTOR (Co)

P1	PM1 PC	PM2 PP	PM3 PS	R1 GR	R2	R3	S1 SR	S2	S3	SS	V	V2	A^2	r_{11}	Validity Nv	Pil.	Bom.	Nav.
....	-09	01	....	10	....	....	13	....	03	....	....	-05	39	77	......	....	....	....
....	-03	-07	....	03	....	....	13	....	03	....	....	09	37	46	......	....	....	....
....	-08	04	....	07	....	....	08	....	01	....	....	-02	33	72	......	....	....	....
....	-01	-03	....	02	....	....	09	....	-04	....	....	[2]*66*	50	56	......	....	....	....

FACTOR (I_1)

P1	PM1 PC	PM2 PP	PM3 PS	R1 GR	R2	R3	S1 SR	S2	S3	SS	V	V2	A^2	r_{11}	Nv	Pil.	Bom.	Nav.
....	11	....	15	41	....	....	06	....	....	....	17	01	69	....	2112	[3]21	....	....
....	06	....	20	33	....	....	17	....	....	....	31	-01	76	87	......	....	....	....
....	01	....	02	22	....	....	22	....	....	....	19	04	45	84	1202	23	....	....

FACTOR (I_2)

P1	PM1 PC	PM2 PP	PM3 PS	R1 GR	R2	R3	S1 SR	S2	S3	SS	V	V2	A^2	r_{11}	Nv	Pil.	Bom.	Nav.
....	10	....	*07*	16	....	....	02	....	....	....	26	17	54	75	5163	16	....	....
....	10	....	17	20	....	....	16	....	....	....	23	08	59	89	......	....	....	....

FACTOR (I_3)

P1	PM1 PC	PM2 PP	PM3 PS	R1 GR	R2	R3	S1 SR	S2	S3	SS	V	V2	A^2	r_{11}	Nv	Pil.	Bom.	Nav.
[2]*46*	13	....	-06	18	01	-01	32	....	....	....	17	11	69	73	2279	18	....	....
....	10	....	17	20	....	....	16	....	....	....	23	08	59	89	......	....	....	....
-06	00	....	11	24	....	....	[2]*46*	....	....	....	16	10	64	81	877	[3]17	....	....
....	....	....	....	03	16	32	06	....	....	....	15	04	30	78	2797	00	....	....
12	....	....	....	[2]*46*	05	*38*	26	....	....	....	20	16	72	85	990 1291	[3]11	....	 48

FACTOR (J)

P1	PM1 PC	PM2 PP	PM3 PS	R1 GR	R2	R3	S1 SR	S2	S3	SS	V	V2	A^2	r_{11}	Nv	Pil.	Bom.	Nav.
08	....	....	....	[2]*40*	....	....	10	....	....	....	03	04	51	41	......	....	....	....
[2]*56*	....	....	....	13	....	....	01	....	....	....	17	30	48	62	4760	[3]13	....	....
35	....	....	....	21	....	....	19	....	....	....	[2]*39*	00	59	66	247	00	....	....
....	....	....	....	10	....	....	....	....	....	....	27	08	25	23	746	17	....	....
56	....	....	....	-04	....	....	00	....	....	....	01	07	39	42	1283	[3]14	....	....
17	....	....	....	[2]*30*	15	01	-01	....	....	....	05	19	48	68	682	[3]19	....	....

FACTOR (LE)

P1	PM1 PC	PM2 PP	PM3 PS	R1 GR	R2	R3	S1 SR	S2	S3	SS	V	V2	A^2	r_{11}	Nv	Pil.	Bom.	Nav.
....	....	....	....	14	....	....	01	....	....	....	05	16	42	59	859	18	....	....
....	....	....	....	13	....	....	33	....	....	....	10	28	52	69	2721 608 645	[3]24	.... [3]03	.... [3]28
....	....	....	....	13	....	....	06	-04	....	....	15	03	27	....	3616 844 558	[3]17	.... [3]10	.... [3]13
....	....	....	....	21	....	....	18	06	....	....	-12	16	38	63	4048 605 600	19	.... 00	 18
....	14	....	-06	06	....	....	06	08	....	....	19	[2]*38*	35	72	2732 611	[3]20	....	.... [3]34

Table 28.15

Test and Code No.	N_t	Ca	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_1 VM	M_3	N	P	PI
MATHEMATICAL BACK															
Biographical Data, CE602D (Navigator)	3000							42	−04				19	10	−14
Mathematics A, CI702F	3000							37	07				[2]61	06	00
MECHANICAL EXPERI															
Tool Function, CI906A	153						11		77					30	
Mechanical Information, CI905A	3791		01	−09	−03	00	07		74				−08	00	09
Mechanical Principles, CI903A	7385		13	−05	14	03	−04	04	60		02		−07	08	−02
Mechanical Principles, CI903B	[illegible]	17						−12	58		02		−13	01	
Practical Judgment, CI301BX1 (Mechanical Items)	202					36			54				07	20	
Physical Principles, CI801BX	1900							−11	51					−03	
Biographical Data, CE602D (Pilot)	3000							29	50				−20	14	09
Driving Skill, CI307AX1	202					21			46				17	08	
Judgment (Mechanical) AAFQE JR P-4	1713					00			45						
Mechanical Comprehension, AC10B	153						−09		43					08	
Mechanical Comprehension, AC10D	870						−01		42	01	04	15		13	
Mechanical Functions, CI907A	153					05	05		42					35	
Judgment Items, AC10A, AAFQE JR F-2	1713					17			41						
Technical Vocabulary, CE505C (Pilot)	3000								39				−08	17	[2]34
Information in Judgment, AAFQE JR P-8	1713					13			39						
PAIRED ASSOCIATES															
Memory for Landmarks, CI510AX1	417									61	20	[2]44		16	
Plane-Name Memory, CI506AX1	238									58	02	[2]61		29	
Memory for Plane Silhouettes, CI503AX1	417									56	06	−09		34	
Memory for Ships, CI504AX1	239									50	06	20		29	
VISUAL MEMORY															
Map Memory, CI505AX2	238									14	58	07		35	
Map Memory, CI505AX3	176									18	55			18	
Map Memory, CI505AX1	179									09	54			35	
Map Memory, CI505BX1	238									41	52	05		22	
Plane Formation, CP805B	392						00				50		15	22	
Directional Orientation, CP515B	392						06				38		22	07	
Pattern Analysis, CP512A	392						04				35		26	26	

(Continued)

GROUND FACTOR (MB)

Pl	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	h^2	r_{11}	Validity Nv	Pil.	Bom.	Nav.
....	04	....	....	00	....	....	−09	....	....	05	01	....	25	46	9600 2347	0[illegible]	....	 23
....	−07	....	....	24	....	....	07	....	....	18	[a]37	....	64	....	8840	13	....	....

ENCE FACTOR (ME)

Pl	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	h^2	r_{11}	Nv	Pil.	Bom.	Nav.
....	....	....	....	05	....	....	18	....	....	....	−02	−11	74	....	78	17	....	....
10	03	−01	−11	04	....	....	04	....	....	....	11	15	65	85	3151 1829 2701	28	.... 00	 08
13	09	−02	−10	20	01	00	22	04	....	....	20	[a]61	84	76	10925 1829 2701	37	.... 10	 25
....	18	−02	....	34	....	....	12	....	28	03	03	[a]64	93	....	9930	[a]33	....	....
00	....	....	....	15	....	....	05	....	....	....	12	29	50	62	2487	[a]18	....	....
....	....	....	....	....	....	....	17	....	....	11	38	....	46	....	6070	10	....	....
....	22	....	....	−07	....	....	−05	....	....	−15	−07	....	46	86	7924 1837	30	....	 10
15	....	....	....	11	....	....	15	....	....	....	−05	[a]48	53	55	520	[a]22	....	....
....	....	....	....	12	....	....	....	....	....	....	25	28	37	42	746	30	....	....
....	....	....	....	32	....	....	36	....	....	....	25	27	56	....	7446	24	....	....
....	....	....	....	19	....	....	27	....	....	....	33	34	55	....	6270	22	....	....
....	....	....	....	22	....	24	−03	....	....	....	24	04	41	80	932	[a]29	....	....
....	....	....	....	07	....	....	....	....	....	....	20	13	28	46	746	15	....	....
....	−01	05	....	00	....	....	−01	....	....	....	[a]41	01	47	80	3151 1829 2701	21	.... 00	 11
....	....	....	....	11	....	....	....	....	....	....	19	04	23	36	746	07	....	....

MEMORY FACTOR (M_1)

Pl	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	h^2	r_{11}	Nv	Pil.	Bom.	Nav.
....	....	....	....	15	....	....	07	....	....	....	07	−13	68	82	1421	[a]10	....	....
....	....	....	....	10	....	....	06	....	....	....	13	....	71	82	1771 1652	22	....	.... [a]15
....	....	....	....	−06	....	....	38	....	....	....	16	02	61	62	873	21	....	....
....	....	....	....	10	....	....	31	....	....	....	−08	....	51	66	1694	17	....	....

FACTOR (M_s)

Pl	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	h^2	r_{11}	Nv	Pil.	Bom.	Nav.
....	....	....	....	17	....	....	16	....	....	....	23	....	80	83	1312 1579	17	....	 34
....	....	....	....	14	....	....	21	....	....	....	31	31	61	67	176	16	....	....
....	....	....	....	06	....	....	16	....	....	....	42	26	70	84	793	17	....	....
....	....	....	....	23	....	....	06	....	....	....	05	....	56	66	1063	19	....	....
....	....	....	....	22	....	....	16	02	....	....	....	16	44	84	936	12	....	....
....	....	....	....	31	....	....	[a]41	06	....	....	....	26	56	74	931 751	26	....	 38
....	....	....	....	15	....	....	15	12	....	....	....	18	35	87	640	16	....	....

Table 28.15

MEMORY III

Test and Code No.	N_t	C_s	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Plane-Name Memory, C1506AX1	[illegible]	...	...	...	...	...	...	...	...	[2]68	02	51	...	29	...
Memory for Landmarks, C1510AX1	417	...	...	...	...	...	...	...	...	[2]61	20	44	...	16	...

NUMERICAL

Test and Code No.	N_t	C_s	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Numerical Operations, C1701B (Back)	6266	...	07	06	04	...	−03	−05	−03	...	...	...	81	04	08
Numerical Operations, C1701B (Front)	6266	04	01	02	−00	...	11	−06	−09	...	...	...	78	10	02
Plotting Accuracy, CE453A (Total Rights)	354	19	...	...	...	...	...	...	...	...	...	...	69	...	...
Numerical Operations, C1701B (Total)	372	...	...	...	02	14	...	...	−06	...	...	...	66	06	...
Mathematics B, C1206B, C1706A	3066	...	...	...	07	04	04	...	08	...	13	...	57	−01	−02
Dial and Table Reading, CP621A, CP622A	6000	...	...	...	...	...	...	15	−02	...	...	...	53	31	11
Complex Scale Reading, CE454A (Total R)	354	05	...	...	...	...	...	...	...	...	...	...	52	...	...
Mathematics A, C1702F	3000	...	...	...	...	...	...	[2]37	07	...	...	...	51	06	00
Plotting Test, CE452A (Total Right)	354	22	...	...	...	...	...	...	...	...	...	...	51	...	...
Dial Reading, CP622A	302	...	...	...	...	...	...	...	...	...	06	...	50	27	...
Table Reading, CP621A	302	...	...	...	...	...	09	...	...	...	10	...	50	30	...
Mathematics B, C1206C	3266	13	04	00	16	...	−01	12	15	...	...	...	48	−01	−07
Number Series Completion, C1215AX1	302	...	...	...	20	...	...	...	...	...	...	...	47	05	...
Directional Plotting, CE455A (Total R)	354	−03	...	...	...	...	...	...	...	...	...	...	44	...	...
Organizational Planning, C1407AX1	302	...	...	...	...	13	...	...	20	...	...	...	41	18	...
Organizational Planning, C1407BX1	266	...	02	35	28	...	03	...	20	...	...	...	38	16	...

PERCEPTUAL SPEED

Test and Code No.	N_t	C_s	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Speed of Identification, CP610A (Non-Rotated)	302	...	...	...	...	...	...	...	...	...	−14	...	−06	70	...
Speed of Identification, CP610A (Rotated)	7649	00	04	02	12	10	03	01	09	05	09	09	04	64	06
Spatial Orientation I, CP501A	153	...	...	...	...	...	02	...	02	...	...	...	...	62	...
Spatial Orientation I, CP501B	7802	...	−02	12	17	03	13	05	01	12	20	00	18	62	16
Spatial Orientation II, CP503B	6409	...	...	...	...	...	...	06	13	09	06	06	−03	54	09
Pursuit-Path Tracing, CP312A	796	...	...	...	17	−01	...	...	07	...	11	...	23	51	...
Map Planning, C1412AX1	170	...	...	...	−01	16	...	...	−08	...	...	...	12	45	...
Block Counting, CP812A	303	...	...	...	...	...	04	...	...	...	38	...	−12	43	...
Planning a Circuit, C1401A	302	...	...	...	...	19	...	...	24	...	...	...	02	41	...

(Continued)

FACTOR (M_2)

P1	PM1 PC	PM2 PP	PM3 PS	R1 GR	R2	R3	S1 SR	S2	S3	SS	V	V2	h²	r11	Validity Nv	Pil.	Bom.	Nav.
....	....	....	....	10	....	....	06	....	....	....	13	....	71	82	1771 1652	22	....	 21
....	....	....	....	15	....	....	07	....	....	....	07	−13	68	82	1421	[3]10	....	....

FACTOR (N)

....	03	14	10	11	....	....	05	....	....	....	06	05	73	79	3151 1829 2701	07	.... 12	 41
....	06	22	23	02	....	....	−01	....	−06	....	05	−10	78	84	3151 1829 2701	−01	.... 12	 38
....	15	18	....	02	....	....	20	....	38	....	....	03	75	78	....	....	....	....
03	....	....	....	16	11	07	07	....	....	....	00	−06	53	81	....	....	....	....
10	−03	04	....	40	−04	00	13	−04	....	....	29	22	68	80	3151 1829 2701	10	.... 13	 49
....	02	14	....	16	....	....	[3]46	....	....	−02	10	07	65	87	10928 1829 2701	22	.... 18	 55
....	−02	17	....	19	....	....	33	....	32	....	....	−02	55	86	....	....	....	....
....	07	....	....	34	....	....	07	....	....	16	[3]37	....	64	....	8540	13	....	....
....	15	20	....	00	....	....	25	....	[3]46	....	....	09	65	80	....	....	....	....
....	....	....	....	24	....	....	[3]41	....	....	....	....	−13	50	76	8630 1048	16	 21	....
....	....	....	....	13	....	....	35	−11	....	....	....	−14	53	84	9602 1021	18	 24	....
....	−12	−06	04	[3]47	....	....	18	....	02	14	27	19	72	84	10857	10	....	....
....	....	....	....	38	19	00	04	....	....	....	10	16	47	75	2300	[3]12	....	....
....	−06	26	....	17	....	....	30	....	[3]46	....	....	[3]46	76	76	....	....	....	....
05	....	....	....	29	....	....	29	....	....	....	27	02	80	65	100	32	....	....
....	00	....	06	00	....	....	15	....	....	....	14	−09	46	....	....	....	....	....

FACTOR (P)

....	....	....	....	−06	....	....	10	....	....	....	....	03	83	....	....	....	....	....
03	07	−11	05	−01	31	03	06	−01	13	....	05	10	63	76	10925 1829 2701	18	.... 06	 15
....	....	....	....	12	....	....	13	....	....	....	16	24	50	91	3963 412	17	....	 16
06	07	16	21	12	06	04	10	....	....	00	06	06	60	63	10923 1829 2701	30	.... 13	 33
....	−01	03	....	17	....	....	16	....	....	30	14	25	63	60	10925 1829 2701	28	.... 06	 30
27	....	....	....	−11	−06	06	17	....	....	....	10	−03	50	66	1168	09	....	....
24	....	....	....	31	....	....	27	....	....	....	−06	34	57	78	1530	21	....	....
....	....	....	....	20	....	....	28	21	....	....	....	11	61	....	640	17	....	....
[3]40	....	....	....	10	....	....	38	....	....	....	34	06	57	88	3478	[3]28	....	....

TABLE 28.15

Test and Code No.	N_t	Co	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
PERCEPTUAL SPEED															
Aerial Photographs, QI901A-IV	302	...	...	...	...	...	06	...	...	...	15	...	27	39	...
Decoding, CI214AX1	202	...	...	...	00	...	...	...	...	...	...	...	12	36	...
Decoding, CI214AX2	1000	...	...	...	...	...	...	−06	−17	...	...	...	...	31	...
PILOT INTEREST															
General Information, CE505D (Pilot)	3000	...	...	...	...	...	...	00	30	...	...	...	−10	23	38
Technical Vocabulary, CE505C (Pilot)	3000	...	...	...	...	...	...	...	[2]39	...	...	...	−08	17	34
Technical Vocabulary, CE505C (Bombardier)	3000	...	...	...	...	...	...	...	15	...	...	...	04	08	33
PLANNING															
Planning Air Maneuvers, CI408AX1	202	...	...	...	...	07	...	...	17	...	...	...	16	00	...
Route Planning, CI411AX1	436	...	17	15	37	11	−02	...	06	...	...	...	02	17	...
Planning Air Maneuvers, CI408AX3	633	...	01	06	[2]43	−11	−19	...	20	...	...	...	08	06	...
Planning a Circuit, CI401A	202	...	...	...	...	19	...	...	24	...	...	...	02	[2]41	...
Practical Estimations I, CI308AX1	170	...	...	...	01	[2]36	...	...	33	...	...	...	13	−05	...
Practical Judgment, CI301BX1 (Non-Mechanical)	202	...	...	...	...	[2]39	...	...	13	...	...	...	15	16	...
PSYCHOMOTOR COORDI															
Rotary Pursuit, CM803A-2	6000	−08	...	...	...	...	...	01	09	...	...	...	−02	10	10
Rudder Control, CM120B (Fernald)	354	08	...	...	...	...	...	01	01	...	...	...	−03	−03	...
Complex Coordination, CM701A	7802	06	06	02	08	11	05	12	16	04	07	07	05	20	17
Aiming Stress, CE211A	6000	...	...	...	...	...	...	00	01	...	...	...	−03	08	−04
Finger Dexterity, CM116A	6000	02	...	...	...	...	...	−04	03	...	...	...	10	20	03
PSYCHOMOTOR PRECI															
Discrimination Reaction Time, CP611D	6000	03	...	...	...	...	...	06	02	...	...	...	18	22	08
Finger Dexterity, CM116A	6000	02	...	...	...	...	...	−04	03	...	...	...	10	20	03
PSYCHOMOTOR SPEED															
Log Book Accuracy, X1	266	...	06	06	09	...	02	...	−06	...	...	...	32	13	...
Marking Accuracy, X1	264	...	−07	01	−07	...	−10	...	−04	...	...	...	02	33	...

(*Continued*)

FACTOR (*Continued*)

P1	PM₁ PC	PM₂ PP	PM₃ PS	R₁ GR	R₂	R₃	S₁ Sn	S₂	S₃	SS	V	V₂	h^2	r_u	Validity Nv	Pil.	Bom.	Nav.
....	....	....	....	05	....	....	24	−06	....	....	....	18	35	....	508	18	....	....
....	....	....	....	[3][illegible]	30	[3]37	19	....	....	....	01	01	54	72	1529	[3]12	....	....
....	....	....	....	....	....	....	22	....	....	10	14	....	26	....	1529	[3]12	....	....

FACTOR (PI)

P1	PM₁ PC	PM₂ PP	PM₃ PS	R₁ GR	R₂	R₃	S₁ Sn	S₂	S₃	SS	V	V₂	h^2	r_u	Nv	Pil.	Bom.	Nav.
....	08	....	....	−05	....	....	−04	....	....	....	[3]36	....	43	87	12042	26	....	....
....	−01	05	....	00	....	....	−01	....	....	....	[3]41	01	47	80	3151	21	....	....
															1829	....	00	....
															2701	....	....	11
....	00	12	....	01	....	....	−05	....	....	....	[3]44	15	37	37	3151	12	....	....
															1829	....	04	....
															2701	....	....	15

FACTOR (P1)

P1	PM₁ PC	PM₂ PP	PM₃ PS	R₁ GR	R₂	R₃	S₁ Sn	S₂	S₃	SS	V	V₂	h^2	r_u	Nv	Pil.	Bom.	Nav.
51	....	....	....	30	....	....	23	....	....	....	26	10	54	51	374	[3]27	....	....
47	−10	....	−03	22	....	....	17	....	....	....	−04	29	63	77	1983	[3]16	....	....
46	13	....	−06	18	01	−01	22	....	....	....	17	11	60	73	2270	18	....	....
40	....	....	....	10	....	....	28	....	....	....	24	06	57	88	3578	[3]26	....	....
36	....	....	....	−04	....	....	00	....	....	....	01	07	39	42	1283	[3]14	....	....
36	....	....	....	13	....	....	01	....	....	....	17	30	48	62	4760	[3]13	....	....

NATION FACTOR (PM₁)

P1	PM₁ PC	PM₂ PP	PM₃ PS	R₁ GR	R₂	R₃	S₁ Sn	S₂	S₃	SS	V	V₂	h^2	r_u	Nv	Pil.	Bom.	Nav.
....	54	20	....	00	....	....	10	....	[3]65	....	01	20	70	92	10925	25	....	....
															1829	....	14	....
															2701	....	....	04
....	48	00	....	−03	....	....	13	....	09	−16	09	27	37	....	10345	34	....	....
10	40	10	00	06	06	14	[3]49	−03	14	01	02	16	64	91	10925	20	....	....
															1829	....	13	....
															2701	....	....	27
....	37	00	....	10	....	....	06	....	....	....	06	05	16	76	10925	11	....	....
															1829	....	00	....
															2701	....	....	00
....	34	[3]83	....	00	....	....	12	....	−03	−08	02	−03	27	92	10925	13	....	....
															1829	....	15	....
															2701	....	....	16

SION FACTOR (PM₂)

P1	PM₁ PC	PM₂ PP	PM₃ PS	R₁ GR	R₂	R₃	S₁ Sn	S₂	S₃	SS	V	V₂	h^2	r_u	Nv	Pil.	Bom.	Nav.
....	12	35	....	11	....	....	[3]48	....	06	07	11	20	48	92	10925	25	....	....
															1829	....	[illegible]	....
															2701	....	....	36
....	[3]54	28	....	00	....	....	12	....	−03	−08	02	−03	27	92	10925	13	....	....
															1829	....	18	....
															2701	....	....	30

FACTOR (PM₃)

P1	PM₁ PC	PM₂ PP	PM₃ PS	R₁ GR	R₂	R₃	S₁ Sn	S₂	S₃	SS	V	V₂	h^2	r_u	Nv	Pil.	Bom.	Nav.
....	02	....	60	03	....	....	19	....	....	....	09	−11	56	76	....	....	....	....
....	−06	....	50	06	....	....	−04	....	....	....	05	−03	41	88	....	....	....	....

TABLE 28.15

GENERAL REASONING

Test and Code No.	N_t	Ca	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Mathematics B, CI206B	570	....	....	....	....	....	18	....	−04	09	16	14	....	00	....
Forced Landings, CI652AX4	264	....	16	38	11	....	−08	....	03	....	....	....	−02	07	....
Reasoning (Arithmetic) AAFQE JR P-7	1713	....	....	....	....	08	....	....	21	....	....	....	....	....	....
Mathematics B, CI206C	3266	13	04	00	16	....	−01	12	15	....	....	....	[2]48	−01	−07
Pattern Comprehension, CP503AX1	163	....	....	....	....	....	00	....	06	....	....	....	....	29	....
Spatial Reasoning, CI211BX1	404	....	....	....	[2]38	14	....	....	02	....	....	....	18	16	....
Spatial Visualization II, CI203AX1	202	....	....	....	17	....	....	....	....	....	....	....	17	08	....
Picture Integration, CP104A	392	....	....	....	....	....	−08	....	....	....	28	....	16	27	....
Judgment (Reasoning) AAFQE JR P-8	1713	....	....	....	....	30	....	....	04	....	....	....	....	....	....
Practical Judgment II, CI301BX2 (Work Plan)	170	....	....	....	01	[2]46	....	....	32	....	....	....	15	−05	....
Competitive Planning, CI400AX2	372	....	....	....	33	[2]36	....	....	−02	....	....	....	15	05	....
Decoding, CI214AX1	202	....	....	....	00	....	....	....	....	....	....	....	12	[2]36	....

REASONING II

Test and Code No.	N_t	Ca	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Figure Analogies, CI212AX1	468	....	08	04	34	....	11	....	−05	....	....	....	20	17	....
Gottschaldt Figures, QP901A-III	594	....	....	....	08	....	02	....	....	....	15	....	22	16	....

REASONING III

Test and Code No.	N_t	Ca	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Spatial Reasoning, CI211BX1	404	....	....	....	[2]38	14	....	....	02	....	....	....	18	16	....
Decoding, CI214AX1	202	....	....	....	00	....	....	....	....	....	....	....	12	[2]36	....

SPATIAL RELATIONS

Test and Code No.	N_t	Ca	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Instrument Comprehension, II, CI616B	468	....	11	11	−02	....	13	03	03	....	....	....	09	17	....
Complex Coordination, CM701A	7502	06	06	02	08	11	05	12	16	04	07	07	05	20	17
Planning a Course, CI406AX2	436	....	06	17	[2]41	10	16	....	00	....	....	....	30	05	....
Instrument Comprehension I, CI615A	468	....	−07	21	08	....	02	....	07	....	....	....	10	18	....
Flags, Figures, Cards, CP312A	392	....	....	....	....	....	08	....	....	....	15	....	−05	31	....
Dial and Table Reading, CP621A, CP622A	6000	....	....	....	....	....	....	15	−02	....	....	....	[2]55	31	11
Discrimination Reaction Time, CP611D	6000	03	....	....	....	....	....	08	02	....	....	....	18	22	06
Cubes, CP312A	655	....	06	00	03	....	−13	....	−10	....	−03	....	18	31	....
Dial Reading, CP622A	392	....	....	....	....	....	....	....	....	....	08	....	[2]60	27	....
Directional Orientation, CP515B	392	....	....	....	....	....	06	....	....	....	[2]36	....	22	07	....
Two-Hand Coordination, CM101A	6266	−07	−05	−06	00	....	06	09	40	....	....	....	00	09	08
Picture Integration, CP104A	392	....	....	....	....	....	−08	....	....	....	28	....	16	27	....
Speed Estimation II, CP205A	392	....	....	....	....	....	29	....	....	....	03	....	−08	08	....
Decoding, CI214AX2	1900	....	....	....	....	....	....	−06	−17	....	....	....	....	[2]31	....

(*Continued*)

FACTOR (R_1)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_4	h^2	r_{tt}	Validity Nv	Pil.	Bom.	Nav
....	....	....	....	57	....	....	12	....	....	....	29	10	51	....	12057 1378	09	 13	....
....	07	....	02	53	....	....	14	....	....	....	18	−05	53	....	1310	[a]08	....	....
....	....	....	....	51	....	....	....	....	....	....	37	14	47	59	746	16	....	....
....	−13	−00	04	[a]47	....	....	18	....	02	14	27	19	72	84	16657	10	....	....
....	....	....	....	45	....	....	−01	....	....	....	21	28	41	....	523	09	....	....
12	....	....	....	46	05	[a]58	26	....	....	....	20	16	72	85	909 1201	[a]11	....	48
....	....	....	....	44	35	36	17	....	....	....	18	[a]48	76	87	2088	17	....	....
....	....	....	....	42	....	....	[a]40	−09	....	....	....	26	61	77	740	23	....	....
....	....	....	....	41	....	....	....	....	....	....	28	14	37	26	746	18	....	....
08	....	....	....	40	....	....	10	....	....	....	03	04	51	41	....	....	....	....
17	....	....	....	36 36	15 30	01 [a]57	−01 19	....	....	....	05 01	19 01	48 54	68 72	682 1529	[a]19 13	....	....

FACTOR (R_2)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_4	h^2	r_{tt}	Nv	Pil.	Bom.	Nav
....	17	....	07	34	40	31	14	....	....	....	23	28	76	83	1009 1675	19	....	61
....	....	....	....	16	39	12	04	−02	....	....	−05	38	43	....	792	24	....	....

FACTOR (R_3)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_4	h^2	r_{tt}	Nv	Pil.	Bom.	Nav
12	....	....	....	[a]46 [a]35	05 30	38 37	26 19	....	....	....	20 01	16 01	72 54	85 72	909 1201 1529	[a]11 13	....	48

FACTOR (S_1)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_4	h^2	r_{tt}	Nv	Pil.	Bom.	Nav
....	−07	....	14	09	26	04	53	....	....	02	24	25	65	84	5880	[a]23	....	....
10	[a]40	10	00	06	04	14	49	−03	14	01	02	16	64	91	10025 1829 2701	30	.... 13	 27
−08	00	....	[illegible]	24	....	....	45	....	....	....	16	10	64	81	877	[a]17	....	....
....	−08	....	24	21	34	00	44	....	....	....	27	07	57	68	9284	[a]20	....	....
....	....	....	....	16	....	....	43	42	....	....	....	14	54	65	1950	[a]24	....	....
....	02	14	....	16	....	....	42	....	....	−02	12	07	65	87	10025 1829 2701	22	.... 18	 64
....	12	[a]56	....	21	....	....	42	....	08	07	11	20	48	92	10025 1829 2701	28	.... 26	 30
....	10	....	−08	38 24	....	....	41 41	25	....	....	14	20 −13	53 59	68 76	2155 8620 1044	[a]23 16	 21	....
....	....	....	....	31	....	....	41	08	....	....	....	28	86	74	931 751	20	....	36
....	34	10	04	−04	....	....	41	....	[a]68	−01	−04	12	85	93	10025 1829 2701	35	.... 12	 19
....	....	....	....	[a]48	....	....	40	−09	....	....	....	30	61	77	740	23	....	....
....	....	....	....	−17	....	....	33 32	02	....	 10	 14	09	28 36	53	1017 1529 250	08 13	....	 24

Table 28.15

Test and Code No.	N_t	Ca	I_1	I_2	I_6	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
SPATIAL II															
Hands, CP512A	658	...	13	-02	06	...	01	...	13	...	-05	...	17	13	...
Flags, Figures, Cards, CP512A	392	...	...	...	...	...	06	...	...	...	15	...	-05	31	...
SPATIAL III															
Two-Hand Coordination, CM101A	6266	-07	-08	-06	00	...	08	09	40	...	...	...	00	09	08
Rotary Pursuit, CM803A-2	6000	-08	...	...	...	...	...	01	06	...	...	...	-02	10	10
Plotting Test, CE452A (Total Rights)	354	22	...	...	...	...	...	...	...	...	...	...	[2]51	...	...
Directional Plotting, CE455A (Total Rights)	354	-03	...	...	...	...	...	...	...	...	...	...	[2]44	...	...
SOCIAL SCIENCE BACK															
Geography	1900	...	...	...	...	...	...	01	12	...	...	...	...	15	...
History, AS153	1900	...	...	...	...	...	...	08	-08	...	...	...	...	08	...
VERBAL															
Technical Vocabulary, CE505C (Navigator)	3638	...	03	03	01	06	14	...	10	...	...	...	14	10	00
Vocabulary, CI604B (AAF)	1900	...	...	...	...	...	...	04	-11	...	...	...	...	11	...
Physics, CI801A	153	...	...	...	...	...	01	...	21	...	...	...	...	17	...
Reading Comprehension, AC10B	202	...	...	...	...	13	...	...	39	...	...	...	20	-02	...
Reading Comprehension, AC10D	570	...	...	...	...	...	-01	...	05	20	18	-03	...	08	...
History, AS153	1900	...	...	...	...	...	...	08	-08	...	...	...	...	08	...
Geography	1900	...	...	...	...	...	...	01	12	...	...	...	...	15	...
General Information, CE505D (Navigator)	3000	...	...	...	...	...	...	27	02	...	...	...	20	10	20
Reading Comprehension, CI614G	6372	...	...	...	07	04	...	05	37	...	...	...	12	03	04
Reading Comprehension, CI614H	266	10	16	26	22	...	-01	05	04	...	...	...	12	-03	...
Vocabulary, AAFQE JR P-1	1713	...	...	...	...	06	...	...	10	...	...	...	...	...	...
Memory for Tactical Plans, CI509AX	179	...	...	...	...	...	...	...	...	10	-12	...	...	-02	...
Reasoning (In Reading), AAFQE JR P-10	1713	...	...	...	...	13	...	...	06	...	...	...	...	...	...
Mathematics A, CI702E	3000	...	...	...	...	...	...	...	-04	...	...	...	42	07	-12
Technical Vocabulary, CE505C (Bombardier)	3000	...	...	...	...	...	...	...	15	...	...	...	04	09	[2]35
Technical Vocabulary, CE505C (Pilot)	3000	...	...	...	...	...	...	...	[2]39	...	...	...	-08	17	[2]34
Sequence of Maneuvers, CI410A	202	...	...	...	...	[2]38	...	...	00	...	...	...	30	00	...
Mathematics A, CI702F	3000	...	...	...	...	...	...	[2]37	07	...	...	...	[2]51	06	00
Syllogisms, AAFQE JR P-11	1713	...	...	...	...	10	...	...	07	...	...	...	...	...	...
General Information, CE505D (Pilot)	3000	...	...	...	...	...	...	00	30	...	...	...	-10	23	[2]38
Reasoning (Deductive) AAFQE JR P-6	1713	...	...	...	...	08	...	...	05	...	...	...	...	...	...

(Continued)

FACTOR (S_2)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	h^2	r_{11}	Validity Nv	Pil.	Bom.	Nav.
....	−07	....	06	02	....	....	17	46	....	....	07	10	35	95	1967	[2]20	....	....
....	....	....	....	16	....	....	[2]43	42	....	....	....	14	54	55	1950	[2]24	....	....

FACTOR (S_3)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	h^2	r_{11}	Validity Nv	Pil.	Bom.	Nav.
....	3/	10	04	−04	....	....	[2]41	....	58	−01	−04	12	85	93	10925 1829 2701	33	.... 12	 19
....	[2]54	20	....	00	....	....	10	....	53	....	01	20	70	92	10925 1829 2701	25	.... 14	 04
....	15	20	....	00	....	....	25	....	46	....	....	09	65	80	....	....	....	....
....	−06	26	....	17	....	....	30	....	42	....	....	[2]45	76	76	....	....	....	....

GROUND FACTOR (SS)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	h^2	r_{11}	Validity Nv	Pil.	Bom.	Nav.
....	....	....	....	....	....	....	06	....	....	58	[2]61	....	75	....	4088	03	....	....
....	....	....	....	....	....	....	−02	....	....	52	[2]63	....	69	....	5014	−09	....	....

FACTOR (V)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	h^2	r_{11}	Validity Nv	Pil.	Bom.	Nav.
−05	05	01	03	07	18	16	07	....	....	....	75	14	72	77	3151 1829 2701	09	.... 05	 33
....	....	....	....	....	....	....	04	....	....	−01	71	....	53	....	1662	−08	....	....
....	....	....	....	17	....	....	03	....	....	....	68	25	63	66	6079	[2]10	....	....
17	....	....	....	23	....	....	01	....	....	....	65	20	75	....	7273	[2]13	....	....
....	....	....	....	34	....	....	17	....	....	....	64	02	64	....	6270	12	....	....
....	....	....	....	....	....	....	−02	....	....	[2]52	63	....	69	....	5014	−09	....	....
....	....	....	....	....	....	....	06	....	....	[2]58	61	....	75	....	4088	03	....	....
....	−07	....	....	13	....	....	−02	....	....	....	60	....	55	....	12043	09	....	....
18	−02	10	....	27	27	03	15	....	....	....	60	24	79	76	10925 1829 2701	20	.... 13	 38
....	00	16	07	19	....	....	−02	....	06	10	60	30	68	85	4821	17	....	....
....	....	....	....	05	....	....	....	....	....	....	58	06	36	40	746	00	....	....
....	....	....	....	09	....	....	08	....	....	....	57	32	47	68	1564	[2]19	....	....
....	....	....	....	03	....	....	....	....	....	....	57	08	38	50	746	12	....	....
....	00	01	....	12	....	....	16	....	....	....	53	33	63	92	3151 1829 2701	12	.... 14	 42
....	00	12	....	01	....	....	−05	....	....	....	44	15	37	37	3151 1829 2701	12	.... 04	 16
....	−01	05	....	00	....	....	−01	....	....	....	41	01	47	80	3151 1829 2701	21	.... 00	 11
35	....	....	....	21	....	....	19	....	....	....	39	00	59	66	247	00	....	....
....	−07	....	....	24	....	....	07	....	....	16	37	....	64	....	8840	13	....	....
....	....	....	....	06	....	....	....	....	....	....	36	10	20	28	746	08	....	....
....	08	....	....	−05	....	....	−04	....	....	....	35	....	43	87	12043	24	....	....
....	....	....	....	29	....	....	....	....	....	....	34	23	26	35	746	16	....	....

TABLE 28.15

VISUALIZATION

Test and Code No.	N_f	C_a	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Directional Plotting, CE455A (Total Wrongs)	354	[2]*41*	....	....	....	....	....	....	....	....	....	....	08	....	....
Spatial Visualization I, CI204AX1	202	...	...	....	18	....	....	....	....	....	....	....	03	14	....
Mechanical Principles, CI903B	354	1	....	....	....	....	....	−12	[2]*58*	....	02	....	−13	01	....
Spatial Visualization I, CI204AX2	266	....	01	03	34	....	06	....	04	....	....	....	05	20	....
Mechanical Movements, CI904A	153	....	....	....	....	....	−04	....	38	....	....	....	....	37	....
Mechanical Principles, CI903A	7385	....	13	−05	14	03	−04	04	[2]*60*	....	02	....	−07	08	−02
Pattern Comprehension, CP803A	202	....	....	....	13	....	....	....	....	....	....	....	−04	24	....
Directional Plotting, CE455A (Total Rights)	354	−03	....	....	....	....	....	....	....	....	....	....	[2]*44*	....	....
Mechanical Comprehension AAFQE JR P-9	1713	....	....	....	....	05	....	....	39	....	....	....	....	....	....
Mechanical Movements, AAFQE JR P-12	1713	....	....	....	....	09	....	....	22	....	....	....	....	....	....
Driving Skill, CI307AX1	202	....	....	....	....	21	....	....	[2]*46*	....	....	....	17	08	....
Spatial Visualization II, CI203AX1	202	....	....	....	17	....	....	....	....	....	....	....	17	08	....
Map Distance, CP626B	658	....	01	....	−09	....	[2]*50*	....	17	....	04	....	09	01	....

[1]Decimal points omitted.
[2]Loadings in italics indicate that a test is also listed under another factor in which this loading is high.
[3]Derived from combinations of data on similar forms.

(Continued)

FACTOR (Vs)

P1	PM1 PC	PM2 PP	PM3 PS	R1 GR	R2	R3	S1 SR	S2	S3	SS	V	Vs	A²	r11	Validity: Nv	Validity: Pil.	Validity: Rom.	Validity: Nav.
....	−01	−03	....	02	....	....	09	....	−04	....	....	56	50	56	....	....	.....	.....
....	....	....	....	34	32	21	24	....	....	....	15	58	71	35	196	26	.....	.....
....	18	−02	....	34	....	....	12	....	28	03	03	54	93	....	9930[2]	33	.....	.....
....	−04	....	05	39	....	....	14	....	....	....	23	53	69	....	1550 735	12	....	 45
....	....	....	....	25	....	....	28	....	....	....	05	51	59	75	2572	[2]23	.....	.....
13	09	−02	−10	20	01	00	22	04	....	....	20	51	84	76	10925 1829 2701	37	..... 10	 25
....	....	....	....	33	24	13	06	....	....	....	18	50	55	....	1031	16	.....	.....
....	−06	26	....	17	....	....	30	....	[2]42	....	....	45	76	76	....	....	.....	.....
....	....	....	....	18	....	....	....	....	....	....	13	44	46	69	746	32	.....	.....
....	....	....	....	10	....	....	....	....	....	....	05	43	31	65	746	14	.....	.....
15	....	....	....	11	....	....	15	....	....	....	−05	42	53	55	520	[2]22	.....	.....
....	....	....	....	[2]44	35	36	17	....	....	....	18	42	75	87	3098	17	.....	.....
....	14	....	−06	06	....	....	06	08	....	....	19	38	35	72	2752 911	[2]20	....	 34

703326—47—54

PI—The pilot-interest factor is common to the pilot criterion and to tests designed to measure pilot interest.

Pl—This is a planning factor of some type, so called because it is in common to certain planning tests. It may involve visualization of a creative type.

PM_1 (PC)—Psychomotor coordination. This factor is substantial in all psychomotor tests analyzed, with the possible exception of Rudder Control. Whether it represents eye-hand coordination or integration of muscular movements, or both, is not known. It seems best described as general muscular agility.

PM_2 (PP)—This is the psychomotor-precision factor, heavily weighted in the bombardier criterion and in psychomotor tests that require precise manipulations under speed requirements.

PM_3 (PS)—This is the psychomotor-speed factor restricted to two highly similar tests that require simple, rapid movements (marking an answer sheet).

R_1 (GR)—This is the general-reasoning factor found extensively in most reasoning tests and strongly in Arithmetic Reasoning.

R_2—The second reasoning factor is hard to define. It is quite strong in the figure analogies test and other tests that are factorially so complex that the clues to its identity are difficult to isolate.

R_3—The third reasoning factor is strongest in tests that seem to require sequential reasoning and in which frequently one can arrive at the correct answer by elimination of wrong answers.

S_1 (SR)—This is the spatial-relations factor which seems to involve relating different stimuli to different responses, either stimuli or responses being arranged in spatial order. It is not clear whether the appreciation of spatial arrangement of stimuli or of responses separately is the key to the factor.

S_2—This is a spatial factor restricted to a few tests such as Thurstone's hands and flags tests. An appreciation of right-hand-left-hand discrimination may be an important aspect of the factor.

S_3—The third space factor was found in only one analysis—that of the carefulness battery. The nature of S_3 is still unknown.

SS—The social-science background factor has been boldly generalized from its strong communality in History and Geography examinations.

V—The verbal factor is best epitomized by vocabulary tests or simple verbal-comprehension tests.

Vz—This visualization factor is strongest in tests that present a stimulus either pictorially or verbally, and in which some manipulation or transformation to another visual arrangement is involved.

Validities of the Factors

Reference was made earlier to estimates of validities of factors that emerged in analyses of the classification batteries. Those estimates pertained to less than half the entire list that was just given. It is of utmost importance that we examine the additional factors not already represented in the classification battery, for in them lie clues to unique variances not yet covered. If any of them are also components of job criteria and can be shown to be weighted to any significant degree in those criteria, they may be the basis for increasing the validities of composite aptitude scores to a practical degree. It is therefore desirable to extract from the assembled evidence as much indication as we can, inaccurate though it may be, of the loadings for pilot, bombardier, and navigator criteria in the various new factors. Owing to the lack of validation data for tests for bombardier and navigator training, we shall be restricted to the pilot criterion in this study. The least that this account does is to demonstrate the usefulness of a limited list of common reference factors in describing criteria and in predicting validities for tests that have never been validated directly.

A basic equation.—In estimating the loadings for the pilot criterion, three successive steps or procedures were used. Each battery of tests analyzed (excluding the classification batteries) served as a basis for the first two steps. The basic equation that gives the correlation between two variables as a function of their common-factor loadings, when applied to the special case of pilot validity, reads:

$$r_{ap}=a_1p_1+a_2p_2+a_3p_3+\ldots\ldots+a_np_n$$

in which

r_{ap} = the validity of test A for pilot training,

$a_1, a_2, a_3, \ldots a_n$ = loadings of factors 1, 2, 3, ... n in test A,

and

$p_1, p_2, p_3, \ldots p_n$ = loadings of factors 1, 2, 3, ... n in the pilot criterion.

For each test whose factor loadings and pilot validity are known, there is one such equation, in which p_1, p_2, p_3, etc., are unknown. There are as many equations as there are tests and as many unknowns as there are factors. The solution of these simultaneous equations yields estimates of $p_1, p_2, p_3 \ldots p_n$.

Solution of the estimates.—The first step was to estimate the loadings p_1, p_2, p_3, etc., by a least-square solution. The number of unknowns in some sets of equations was reduced by assuming that the loadings for the verbal, numerical, and general-reasoning factors were zero. In still other instances, all of which will be noted in table 28.16, other loadings, principally for factors ME and PM_1 were assumed and corresponding allowances made for them in the equations. These assumed loadings were taken from the December 1942 battery analysis (see table 28.14). After having thus reduced the number of unknowns, the number of equations was then reduced to the number of unknowns by computing sums of squares of the deviations of loadings in each column from the mean of the column, and the sums of cross products between all possible pairs of columns.[a] These equations were then solved simultaneously by a modified Doolittle method.

The second step consisted in minor adjustments in the solutions just described. It was recognized that not all valid factors were represented in each set of equations. Accordingly, one might expect that the basic equation when applied to incomplete data should frequently yield underestimations of validities, but should rarely give overpredictions. In this step, by trial and error, adjustments, usually downward, were made in the estimated values for p_1, p_2, p_3, etc., until a more reasonable fit, in accordance with the principles just stated, was accomplished.

The results from steps one and two are given in table 28.16. A blank in the table indicates that a factor did not appear in an analysis. Assumed loadings are indicated by a superscript "1." For some factors, only one estimate is available, and for others a number of estimates. The stability of any loading, or lack of it, can be seen in the rows of the table. The concordance of many estimates with previous estimates may also be noted. The agreement is generally reassuring, though considerable variation from sample to sample is obvious. Of the seven factors whose loadings had been previously estimated (see table 28.14), judging from the mean loadings given in the next to the last column, the validities of the mechanical-experience and psychomotor-coordination factors had been somewhat underestimated in the December 1942 analysis, and those for perceptual speed, spatial relations, and visualization, had been somewhat overestimated. In the last column are given estimates rounded to the nearest 0.05, which is taken to be the limit of accuracy probably justified here for most factors.

The third step was a checking method which cut across factor batteries. Assuming the loadings in the last column of table 28.16 to be correct, each factor was taken in turn as the "unknown" or X factor whose validity was to be determined. All tests in table 28.15 in which factor X had a substantial loading were examined in the following manner. No test was included unless its loading in X was equal to or greater than 0.30. Using the fundamental equation as the basis, the part validity of each test (omitting factor X) was estimated and a residual validity which could probably be attributed to factor X was computed. The division of the residual by the loading of factor X in the test gave one estimate of the validity of factor X. There were as many estimates of this validity as there were tests involved. A mean of these estimates was taken to be the best value. In some instances, as with the three integration factors

[a] The columns referred to here are those of the rotated factor matrix such as that in table 28.11.

which occurred combined so frequently in the same tests, more than one unknown was assumed. In these instances, simultaneous equations with two or three unknowns were solved, using the residual validities as before.

The factor loadings in the pilot criterion.—This procedure (step 3) led to verification of most of the previous estimates given in the last column of table 28.16. The revised estimates are presented in table 28.17, with all factors except K (kinesthetic) accounted for. Also presented are the numbers of batteries (n) in which estimates had been made, including the two classification batteries—July 1943 and December 1942—as well as the batteries listed in table 28.16. The checking procedure of step 3 is not counted as one of the observations in deriving n.

Comparison of the two tables will show a number of changes brought about by step 3. The most drastic ones are for Reasoning II and Reasoning III whose loadings were reduced from 0.15 and −0.20 to 0.05 and −0.05 respectively. The loading of 0.22 for R_2 found in the judgment-and-reasoning battery is suspect, because it was not clear that the R_2 in that battery was one factor or a combination of two factors. The loading of −0.21 for R_3 is unlikely. One would not expect a negative loading of any substantial size in any factor, though loadings of −0.05 and −0.10 have been accepted in other factors than R_3. In checking (step 3), the loadings were estimated for R_2 and R_3 simultaneously. The least-square solution (in connection with step 3) gave loadings of 0.15 for R_2 and −0.05 for R_3. Certain residuals indicated the reduction to 0.05 for R_2, however, and verified the loading of −0.05 for R_3.

The validity of 0.00 previously estimated for Integration II, becomes 0.10 in table 25.17. PM_1 loses weight from its unparalleled 0.30 in table 28.16 to a value of 0.20. Length estimation has a loss from 0.20 to a more conservative 0.15. M_1 insists upon having a validity of 0.05 in the final checking and M_3 gains from 0.10 to 0.15. P1, which had two estimates of 0.03 and 0.04 but a third of 0.00, gains a place in the nonzero class with a validity of 0.05. S_3 curiously enough shifts from +0.05 to −0.05. With so few tests as we have to represent it, we can only be puzzled by this result and wait for further confirmation one way or the other. Loadings that remain the same through the checking procedure of step 3 were factors I_1, I_2, J, ME, M_2, N, P, PM_2, R_1, S_1, S_2, V, and Vz. Most loadings are left in the rougher values to the nearest five hundredth. Two of the stronger and better established loadings, however, are given to the nearest one hundredth, namely, for factors ME and S_1.

Negative factor validities.—Negative loadings present something of a question. A negative validity would mean that individuals having a high degree of a factor do less well on the average than individuals who have average or low degrees of it. The ability may actually be regarded as a handicap. On this basis, most of the negative validities in table 28.17 can be successfully rationalized. The factors in which the negative loadings occur for the pilot tend to be abilities involving either tedious, re-

TABLE 28.16.—*Estimations of factor validities for the pilot criterion derived from several batteries analyzed.*

Factor	ANALYSIS												
	Mechanical battery	Memory I	Memory II	Foresight and Planning I	Foresight and Planning II	Judgment and Reasoning II	Nonverbal Reasoning	Integration	Perceptual I	Perceptual II	Carefulness	Means	Rounded estimates
I_1	....	....	....	....	....	....	....	0.25	....	....	....	0.25	0.25
I_2	....	....	....	....	....	....	....	.03	....	....	....	.03	.05
I_3	....	....	....	....	....	....	....	−.11	....	....	....	−.11	−.10
L	....	....	....	−0.11	0.15	0.21	....	....	....	....	....	.08	.10
LE	0.20	....	....	....	....	....	....	....	....	0.19	....	.20	.20
ME	.29	[a]0.25	[a]0.25	.26	.27	.14	[a]0.25	.43	[a]0.25	[a].25	[a]0.25	.28	.30
M_1	....	.00	.00	....	....	....	....	....	....	....	....	.00	.00
M_2	....	.05	.10	....	....	....	....	....	.05	.04	....	.06	.05
M_3	....	....	.12	....	....	....	....	....	....	....	....	.12	.10
P	.20	.17	.16	.19	.24	....	.13	[a].20	.15	.14	.14	.17	.15
PI	....	....	....	.03	.00	....	[a].04	....	....	....	....	.02	.00
PM_1	[a].22	[a].22	[a].22	[a].22	[a].22	[a].22	[a].22	.36	[a].22	[a].22	.24	.30	.30
PM_2	....	....	....	....	....	....	....	....	....	....	.00	.00	.00
R_1	[a].00	[a].00	[a].00	[a].00	[a].00	.00	[a].00	[a].00	.00	[a].00	[a].00	.00	.00
R_2	....	....	....	....	....	.22	.12	....	....	....	....	.17	.15
R_3	....	....	....	....	....	....	−.21	....	....	....	....	−.21	−.20
S_1	.24	.32	.39	.22	.15	[a].34	.38	.29	.30	.27	.24	.28	.30
S_2	....	....	....	....	....	....	....	....	....	.07	....	.07	.05
S_3	....	....	....	....	....	....	....	....	....	....	.07	.07	.05
V	[a].00	−.02	[a].00	[a].00	[a].00	−.08	[a].00	[a].00	[a].00	[a].00	[a].00	−.05	−.05
V_z	.08	.07	....	.24	.15	.23	.21	.19	.25	.25	.23	.19	.20

[a] An assumed value based mainly on the analysis of the December 1942 battery.

TABLE 28.17.—*Final estimates of factor validities for the pilot criterion*

Factor	n[1]	Loading	Factor	n[1]	Loading
I_1	1	0.25	PI	3	0.05
I_2	1	.10	PM_1 (PC)	4	.20
I_3	1	−.10	PM_2 (PP)	2	.00
J	3	.10	PM_3 (RS)	0	(2)
LE	2	.15	R_1 (GR)	4	.00
MB	1	.10	R_2	2	.05
ME	7	.27	R_3	1	−.05
M_1 (PM)	2	.05	S_1 (SR)	12	.32
M_2 (VM)	4	.05	S_2	1	.05
M_3	1	.15	S_3	1	−.05
N	2	.00	SS	0	−.10
P	11	.15	V	4	−.05
PI	2	.25	Vz	12	.20

[1] *n* is the number of batteries in which estimates of the factor validity were made.
[2] No estimate made for lack of validity data.

stricting, or symbolic activity, or combinations thereof. The verbal factor is highly symbolic. I_3 involves attention to many details. R_3 seems to be reasoning where fine distinctions are important. The nature of S_3 is not sufficiently known to justify any opinion, nor is its negative loading to be fully accepted. The social-science factor probably represents a negative interest in symbolic, indoor activities, for pilots have a positive interest in overt, outdoor activities.

Validity of Factor Composites

From the data in table 28.17 we can derive an estimate of the maximum validity to be expected for pilot selection on the basis of factor validities, when factors are optimally weighted in a composite score in proportion to those validities. In this discussion, orthogonality of factors is assumed. In the last classification battery in use in the Air Corps during the year 1945, of the factors listed in table 28.17, only J, MB, ME, P, PI, PM_1, S_1, and Vz have positive loadings. A sum of the squares of their validities yields an estimate of the maximum possible multiple correlation squared (R^2). Considering these factors only, the sum of squares is 0.3603, and from this R is 0.60. This probably coincides fairly well with empirical results, except that there is probably some reduction in the obtained multiple R due to the weighting of irrelevant factors like N and R_1, and the positive weighting of V when it should have a negative weight. These irrelevant variances are introduced by the use of impure tests like Arithmetic Reasoning, Reading Comprehension, and Dial and Table Reading. Other discrepancies would be due to the fact that the MB factor is not weighted in the composite and that the K factor is weighted but not represented in the list considered above. The unique contribution of the Rudder Control test is accordingly missing from all the estimates of the multiple R made here.

It is of interest next to see what pilot validities could be obtained if all valid factors were included in the composite and optimally weighted. Consider first only the factors with positive weights. The estimate of R^2 is then 0.4903, and R is 0.70, which represents a gain of 0.10 in maximum validity. This could be accomplished at the cost of adding a maximum of nine tests to the battery, one for each additional factor. If impure tests were used for this purpose, properly combining these factors, less than nine tests would be needed. If negatively weighted factors were also included, R^2 would become 0.5178 and R would be 0.72. This gain of 0.02 would be at the cost of adding four tests (the verbal factor being already covered). The cost in this case might be too great for the gain obtained. There is also the general question of the policy of weighting any ability negatively for fear it might mean adverse selection for some other criterion in the job. Selection of tests that are to contribute negative weights should always be made with care.

What has just been done for pilot tests and factors could also be done for navigator and bombardier. Owing to the lack of data, it has been impossible to make satisfying estimates for navigator and bombardier criteria except for factors in the classification battery (see table 28.14). Referring to the data of that table, particularly data for the factor-analysis solution, we can see what the December 1942 battery validities could have been. Summing and squaring the columns and extracting the square roots yield estimates of maximum validities of 0.50, 0.74 and 0.63 for bombardier, navigator, and pilot composites respectively. The empirical validities fell somewhat short of these values; in fact, they were in the neighborhood of 0.30, 0.65, and 0.53 respectively. The discrepancies can be attributed to improper weighting and possibly to overestimates of factor validities in some instances.

Job Analysis by Factor Analysis

In chapter 1 it was intimated that job-analysis information would be forthcoming in later chapters based upon factor-analysis results. It is time now to make good that promise and to see how the empirically determined factors agree with the traits derived from direct inspection of air-crew jobs.

Inspection categories versus factor categories.—Of the list of 20 traits used in describing the job of the pilot in table 1.5, very few concepts have their counterparts in the factor list in this chapter, and none of these has identical meaning. Judgment in table 1.5 has quite a different connotation than the J factor in table 28.17. Foresight and planning of table 1.5 has only one factor resembling it—the factor doubtfully called Planning (Pi). Memory breaks down into at least three separate abilities. Comprehension is replaced by the verbal factor which is restricted to word material.

Visualization of flight course may mean nothing more than the visualization factor when that term is applied specifically to aviation by aviation observers, but this is very doubtful. Estimation of speed and distance must indeed be separated, for distance-judgment tests correlate very low with speed-judgment tests. A length-estimation factor was found, and it is possible there is also a speed-estimation factor, but none was uncovered by correlation analyses. Sense of sustentation may be akin to the kinesthetic factor found in the Rudder Control test, but no evidence is available regarding this possible connection. "Division of Attention" did not appear as a distinct factor, but one or more of the integration factors may be identified with it. Orientation almost surely breaks down, but its components are not clear. It is doubtful that what aviation observers call orientation depends to any great extent on the S_1 factor. It may depend to some extent on S_2. A factor concerning compass directions might have been expected, but none has as yet made its appearance. Speed of decision and reaction is a superficial concept. The test designed to measure it—Discrimination Reaction Time—proved to break down into S_1, PM_2 and P factors. Factor PM_3 is one type of speed of reaction, but its loading in the Discrimination Reaction Time test is as yet problematical.

Coordination is an ambiguous term, but in table 1.5 it probably refers to purely motor coordination or psychomotor coordination. If so, the PM_1 factor is very close to it but there is evidence, not given in this volume, that there is also a motor coordination factor at the finger-dexterity level.

From this point on in table 1.5, the traits show less and less correspondence with the empirical factors. One reason is the lack of factor-analysis data in the area of temperament. The only temperament factor to be found listed in table 28.17 is that of pilot interest; all others are apparently abilities with the exceptions of mathematical background, mechanical experience, and the social-science factor, which are apparently experience variables.

Factors in table 28.17 not mentioned in table 1.5 are the integration factors, mathematical background, mechanical experience, numerical, the three reasoning factors, and the social-science factor. Of these only mechanical experience and possibly integration I have substantial validity for the pilot. Even so, these two were missed in the list of table 1.5. Two others had been anticipated in connection with navigation, namely, mathematical background and the numerical factor.

Factor categories as an aid to inspection.—Very few factors, as was said before, were exactly what had been expected. Having brought them to light, however, we can now ask whether future observations of jobs would be facilitated by their use as compared with the trait categories originally employed. It is quite possible that this is so. As psychologists who worked with these factors became more and more familiar with them, it became natural and apparently easy to examine a new test or a new task and to predict with some confidence the factors probably present and their relative importance. It is believed that with increased definition of variables, this type of analysis by inspection will take on greater utility than has heretofore been possible with the use of old concepts. The stability (reproducibility) of the same variables and their communicability lend great assurance that this is true.

THE PREDICTION OF TEST VALIDITIES

One virtue of factor theory which has as yet been mentioned only incidentally is its utility in the prediction of validities of tests. This is possible by means of the basic equation given on page 839, if enough of the factor loadings are known. Assuming that both new test and criterion have been fully accounted for in terms of known factors, the prediction of validity of the test scores can be made. If one or more factors are unaccounted for, the prediction will fall short of the validity to be obtained, but it will at least give a notion of the minimum validity to be expected if all known loadings are positive.

In table 28.18 are given the predicted and obtained pilot validities of all tests that have been analyzed and for the description of which this volume is primarily responsible. Since the factor loadings used for the predictions were derived from most of the same tests, the results are in a large sense merely a verification of the fit of the factor loadings estimated. The discrepancies between predicted and obtained validities indicate the degree of divergence between assumption and fact.*

There should be little concern regarding positive discrepancies, particularly in the case of tests whose communalities fall short of their respective reliabilities. The differences are probably due to unknown common factors. If a discrepancy of this kind is fairly large, the test deserves thorough scrutiny for hypotheses as to the unknown factor or factors and a research project to define better the factor and to maximize its variance in new tests. Tests in this category are Competitive Planning (deficit validity of 0.14), Mechanical Functions (deficit of 0.14), Memory for Tactical Plans (deficit of 0.13), Number Series Completion (deficit of 0.09), and Pursuit (deficit of 0.12). In all of these tests there is plenty of room for common-factor variance not now accounted for. Granted that these deficits are not due to sampling errors, these tests deserve future intensive study.

Large negative discrepancies are disturbing for they show that the obtained validity is more than accounted for by known factors. In one or two instances such discrepancies led to a reexamination of the obtained validities and as a result gross errors were found and corrected. In other instances we may attribute the discrepancies to sampling errors, if N is relatively small. Experience has shown that with samples of 200 pilots, validities fluctuate all the way from 0.20 to 0.75 (when the mean is about 0.50). With large samples, one might well suspect errors in the estimation of factor loadings in the test or in the criterion.

Within the group of tests having serious over-predictions of validity are the following: Judgment of Proportions (excess validity of 0.10), Mechanical Movements (excess of 0.11), Practical Judgment (mechanical) (excess of 0.10), Sequence of Maneuvers (excess of 0.10), and Tool

* A really crucial test, of course, would be to predict validities in advance and compare predictions with validities obtained in new samples. The obtained validities in table 28.18 are the same as those given in table 28.15.

TABLE 28.18.—*Estimated and obtained pilot validities of analyzed tests described in this volume*

Test	Estimated r'	Obtained r_{bis}	Discrepancy $r_{bis} - r'$	N
Biographical Data, CE602D (N)	−0.02	0.02	0.04	9,600
Biographical Data, CE602D (P)	.21	.30	.09	7,924
Block Counting, CP512A	.21	.17	−.04	829
Code Analysis, CI653AX2	.07	....	....	....
Combat Planes, CI655AX5	.19	....	...	....
Competitive Planning, CI409AX2	.05	.19	.14	682
Complex Scale Reading, CE454A (Total R)	.08	....	...	....
Complex Scale Reading, CE454A (Total W)	−.04	...	...	....
Cubes, CP512A	.21	.23	.02	2,155
Decoding, CI214AX1	.11	.13	.02	1,529
Decoding, CI214AX2	.08	.13	.05	1,529
Dial Reading, CP622A	.15	.16	.01	8,630
Table Reading, CP621A	.14	.18	.04	9,602
Dial and Table Reading, CP621A, CP622A	.23	.22	−.01	10,925
Directional Orientation, CP515B	.22	.26	.04	931
Directional Plotting, CE455A (Total R)	.15	....	....	....
Directional Plotting, CE455A (Total W)	−.14	....	....	....
Driving Skill, CI307AX1	.30	.32	.02	520
Figure Analogies, CI212AX1	.15	.19	.04	1,008
Figure Classification, CI213AX1	−.02	.00	.02	2,797
Flags, Figures, Cards, CP512A	.25	.24	−.01	1,950
Flight Formations, CI654AX5	.21	.23	.02	1,302
Following Directions, CP402A	.13	.11	−.02	5,163
Following Oral Directions, CI651AX3	.20	.24	.04	4,797
Forced Landings, CI652AX4	.12	.08	−.04	1,310
General Information, CE505D (Navigator)	.05	.09	.04	12,043
General Information, CE505D (Pilot)	.20	.24	.04	12,043
General Information, CE505C (Technical Vocabulary—Bombardier)	.13	.12	−.01	3,151
General Information, CE505C (Technical Vocabulary—Navigator)	.11	.09	−.02	3,151
General Information, CE505C (Technical Vocabulary—Pilot)	.19	.21	.02	3,151
Geography	−.01	.03	.04	4,988
Hands, CP512A	.16	.20	.04	1,967
History, AS153	−.09	−.09	.00	5,014
Instrument Comprehension I, CI615A	.19	.20	.01	9,284
Instrument Comprehension II, CI616B	.30	.32	.02	8,889
Judgment of Proportions, CP206B	.18	.08	−.10	765
Log Book Accuracy	.06	....	....	....
Map Distance, CP626B	.22	.20	−.02	2,752
Map Memory, CI505AX1	.17	.16	−.01	212
Map Memory, CI505BX1	.11	.19	.08	1,083
Map Memory, CI505AX2	.14	.17	.03	1,312
Map Memory, CI505AX3	.18	.16	−.02	176
Map Planning, CI412AX1	.22	.21	−.01	1,530
Marking Accuracy	−.02	....	....	.....
Mathematics A, CI702E	.07	.12	.05	3,151
Mathematics A, CI702F	.04	.13	.09	8,840
Mathematics B, CI206B	.09	.09	.00	12,057
Mathematics B, CI206C	.07	.10	.03	18,657
Mathematics B, CI206B, CI706A	.09	.10	.01	3,151
Mechanical Functions, CI907A	.15	.29	.14	932
Mechanical Information, CI905A	.28	.28	.00	3,151
Mechanical Movements, CI904A	.34	.23	−.11	2,572
Mechanical Principles, CI903A	.37	.37	.00	10,925
Mechanical Principles, CI903B	.31	.33	.02	9,930
Memory for Landmarks, CI510AX1	.12	.10	−.02	1,421
Memory for Plane-Names, CI506AX1	.16	.22	.06	1,771
Memory for Plane Silhouettes, CI503AX1	.19	.21	.02	873
Memory for Ships, CI504AX1	.20	.17	−.03	1,694
Memory for Tactical Plans, CI509AX	.06	.19	.13	1,564
Nearest Point—Point Distance, CP607B	.22	.19	−.03	4,045
Number Series Completion, CI215AX1	.04	.13	.09	2,309
Numerical Operations, CI701B (Back)	.06	.03	−.03	3,151
Numerical Operations, CI701B (Front)	.01	−.01	−.02	3,151
Numerical Operations, CI701B (Total)	.01	....	....	....
Organizational Planning, CI407AX1	.18	.25	.07	102
Organizational Planning, CI407BX1	.12	....	....	....
Path Length, CP628B	.17	.23	.06	2,491
Pattern Analysis, CP512A	.15	.16	.01	640
Pattern Assembly, CP804A	.16	.18	.02	839
Pattern Comprehension, CP803A	.14	.16	.02	1,160
Pattern Comprehension, CP803AX1	.10	.09	−.01	163
Physics, CI801A	.11	.10	−.01	6,079
Physical Principles, CI801BX	.15	.10	−.05	6,079
Picture Integration, CP104A	.22	.28	.06	740
Planning a Circuit, CI401A	.26	.26	.00	3,578
Planning a Course, CI406AX2	.19	.17	−.02	877
Planning Air Maneuvers, CI408AX1	.16	.27	.11	374
Planning Air Maneuvers, CI408AX3	.15	.18	.03	2,279

Table 28.18.— (Cont'd)

Test	Estimated r'	Obtained r_{bis}	Discrepancy $r_{bis} - r'$	N
Plotting Test, CE452A (Total R)	.10	....	....	....
Plotting Test, CE452A (Total W)	.01	....	....	....
Plotting Accuracy, CE453A (Total R)	.08	....	....	....
Plotting Accuracy, CE453A (Total W)	.00	....	....	....
Practical Estimations I, CI308AX1	.15	.14	−.01	1,283
Practical Estimations II, CI308AX1	.14	.13	−.01	1,283
Practical Judgment (Mechanical Items) CI301BX1	.28	.18	−.10	2,487
Practical Judgment (Non-Mechanical) CI301BX1	.17	.13	−.04	4,760
Practical Judgment I (Non-Mechanical) CI301BX3	.23	....	....	....
Practical Judgment II (Work Planning), CI301BX3	.17	....	....	....
Pursuit-Path Tracing, CP512A	.13	.25	.12	1,238
Reading Comprehension, CI614G	.20	.20	.00	10,925
Reading Comprehension, CI614H	.06	.17	.11	4,821
Route Planning, CI411AX1	.19	.16	−.03	1,983
Sequence of Maneuvers, CI410A	.10	.00	−.10	247
Shorter Line—Line Length, CP606B	.11	.17	.06	3,616
Shorter Path—Path Distance, CP608B	.25	.24	−.01	5,054
Signal Interpretation, CI656AX2	.18	.21	.03	2,112
Spatial Orientation I, CP501A	.18	.17	−.01	3,965
Spatial Orientation I, CP501B	.23	.20	−.03	10,925
Spatial Orientation II, CP503B	.24	.26	.02	10,925
Spatial Reasoning, CI211BX1	.10	.11	.01	999
Spatial Visualization I, CI204AX1	.19	.15	−.04	1,445
Spatial Visualization I, CI204AX2	.15	.12	−.03	1,685
Spatial Visualization II, CI203AX1	.12	.17	.05	3,088
Speed of Identification (Non-Rotated), CP610A	.14	....	....	....
Speed of Identification (Rotated), CP610A	.23	.18	−.05	10,925
Tool Function, CI906A	.31	.17	−.14	78
Vocabulary, CI604B	−.03	−.08	−.05	1,662

Function (excess of 0.14). Two tests in this group have validities based on relatively small samples—Sequence of Maneuvers (N = 247) and Tool Function (N = 78). Other validities are based on large samples. It is notable that with one or two exceptions, the tests in this group have communalities which approach the reliabilities fairly closely, in contrast to tests in the group mentioned in the last paragraph.

The general picture of the goodness of fit of predicted validities is given in table 28.19. Nearly half of the discrepancies are within a range of ±0.02 of zero and about two-thirds are zero or above. The mean discrepancy is only 0.01, which is lower than would have been expected. This fact may indicate that there has been some tendency to overestimate the factor loadings of the pilot criterion. It might also mean that the list of factors mentioned in this chapter goes much further toward covering the pilot criterion than might be supposed. The coefficient of correlation between predicted and obtained validities is 0.81, which indicates considerable agreement.

TABLE 28.19.—*Distribution of discrepancies between predicted and obtained pilot validities in 90 Tests*

Discrepancy	Frequency	Discrepancy	Frequency
+0.13 to +0.17	3	−0.12 to −0.08	4
+.08 to +.12	7	−.17 to −.13	1
+.03 to +.07	21	N	90
−.02 to +.02	41	M	.011
−.07 to −.03	13		

Validities of Non-Validated Tests

The predicted validities of the small number of tests in table 28.18 that have not been validated do not reveal any spectacular results, but they do provide some guesses as to the probable usefulness of certain tests.[10] In general, the carefulness tests do not promise any added validity for the pilot unless the carefulness factor itself proves to be valid. This is regarded as highly unlikely. If anything, one might predict a negative validity of the carefulness factor for pilots, in using the graduation-elimination criterion. Against an accident criterion, the result to be expected is merely an interesting open question. If, as indicated, the carefulness scores are not valid for the pilot, but are valid for the navigator, as predicted by the carefulness hypothesis, a good classification instrument would be available.

The test Combat Planes offers substantial added validity, because of its unique variance in Integration I. The predicted validity of 0.18 depends largely upon the estimated validity of 0.25 for the I_1 factor which in turn is based upon slender evidence.

Another test of special interest in the nonrotated form of Speed of Identification, which has an expected validity of 0.13. This is much lower than that predicted or obtained for the rotated form, but the nonrotated form is purer—its unique contribution to the pilot stanine would be presumably just as large and its irrelevant variance would be practically nil.

CONCLUDING STATEMENT

Advantages of the Factorial Approach

The discussions in this chapter have shown several advantages of the factor-analysis approach in a test-development program. Its favorable features as they appear to the writer may be summarized as follows:

(1) It provides an exact, quantitative picture of tests and criteria in terms of a limited number of stable, meaningful categories.

(2) It enables us to understand why some tests are valid and why others are not.

(3) It makes possible the substitution of one test for another in terms of equivalent factor patterns.[11]

(4) It leads to the discovery and development of pure tests whose contributions are unique. Such tests are demanded for precise classification.

(5) It leads to the discovery of valid factors not known or even suspected before.

[10] It should be remembered that these are minimum predictions of validity, since not all valid factors may be accounted for.

[11] Just before the end of war II large-scale studies were begun at Keesler Field and at Sheppard Field, under the direction of Psychological Research Unit No. 2. Large numbers of tests were assembled into half-day batteries and administered in such order that correlations of each test with all other tests could be obtained on relatively large samples. It was planned to factor analyze these large matrices. (See appendix C.)

(6) It makes possible the prediction of validities of tests and of composites before the event of validation by ordinary methods.

(7) It should in time make possible a much more enlightened and direct job analysis by inspection of jobs, and job description in factor-category terms.

(8) It serves as a guide in new test development, making possible the avoidance of irrelevant variances such as numerical, verbal, or reasoning when not wanted.

(9) It makes possible the compilation of new test batteries to meet new selection needs when new criteria are described in factorial terms.

Most of these features have been brought out and their applications have been illustrated in this and in preceding chapters. Other features will receive further mention and use in the concluding chapter which immediately follows.

CHAPTER TWENTY-NINE

General Conclusions[1]

The preceding chapter brought together many of the threads of discourse concerning test development which were followed in somewhat isolated but systematic fashion in earlier chapters, and in it an attempt was made to obtain a unified picture. The centralizing principle was that of factorial structure by which a large number of the printed tests developed in the Army Air Forces could be aligned in families and related to validity for the selection of pilot, navigator, and bombardier.

Unfortunately, many intellectual and perceptual tests, more recently developed and never analyzed, and almost all of the temperament tests had to be omitted from that significant type of picture. There remains, therefore, the need for a final glance over the contents of earlier chapters in order to see what the positive gains have been; to evaluate what was done; and to note what was left undone. From such a review, investigators either in the Air Forces or elsewhere may be more profitably guided in the next steps.

This chapter will accordingly present first a summarizing paragraph or two concerning each area of test development. Each area will correspond usually with that of an entire chapter on tests. Reference will also be made to the job-analysis concepts that are most closely related to the test area under discussion. Some of the general lessons learned in the development of printed tests will be briefly mentioned, with implications for future research and future application of tests in classification batteries. Other suggestions which more naturally grow out of development of apparatus tests and motion-picture tests, and out of the experiences of classification testing, will be found in other reports.

It might be in order to devote some space to the implications of what has been learned in this program for general industrial and vocational testing. It is believed, however, that this volume should be confined to a descriptive account of a certain segment of experiences gained by aviation psychologists in a particular program. It is also believed that one who reads liberal portions of the volume will not fail to find implications for more general testing programs if he reads with that intention. Some of the general comments made in the later part of this chapter may serve to direct the attention of such a reader.

A REVIEW OF THE TEST AREAS

As each area of tests is mentioned, several considerations will be given attention. The most closely related job-analysis categories will be men-

[1] Written by the editor.

tioned and the bearing that printed-test development and research have had upon those categories will be pointed out. On the other hand, the fundamental traits or factors that characterize the area will be mentioned, with emphasis upon whatever unique factors the area has to offer. The validity of each area or subarea for success in each type of training will be given a summary statement in nonnumerical terms, where the facts are known, and explanations of validity in terms of known factorial composition will be cited. The needs for further test development and research will be pointed out and the kind of research indicated.

Verbal Ability Tests

The list of 20 job-analysis traits for the pilot, which was given most attention in the program, included among the four intellectual traits the category of comprehension. Two types of comprehension tests were developed: verbal and mechanical. The latter were found later to justify a new category of their own, namely, the mechanical group. This left the verbal-comprehension tests as the only candidates for the comprehension category.

The verbal tests fall into two groups: general vocabulary and reading comprehension (see ch. 5). The vocabulary tests proved to be the purest and strongest as measures of the verbal factor. This factor seems to be represented in the navigator criterion only; vocabulary tests have no validity for either bombardier or pilot. In fact, they have a slight negative validity for the pilot, which, taken with other facts, indicates that the verbal factor is correlated slightly negatively with the pilot criterion. This kind of relationship would surely not hold for the lower levels of verbal ability in which one would expect a positive correlation. This reasoning leads to the inference that over the entire IQ range (for most IQ's are predominantly indices of verbal comprehension) the regression of pilot-training success on the verbal factor is curvilinear.

The reading-comprehension tests used in the Air Forces were of complex factorial composition, and as a result, they were valid to some extent for all three air-crew positions. They were valid for the navigator because of their verbal variance (their strongest factor), the general reasoning factor (next strongest), to a smaller extent because of visualization, and even to some extent because of a slight numerical variance. They were valid for the pilot primarily because of a substantial loading in the mechanical-experience factor and to a less extent by reason of some visualization variance. The smallest validity was for the bombardier, but what there was could be accounted for by the visualization and numerical components.

Since all the factorial components of the reading-comprehension tests are better measured by purer and stronger tests for them, it is strongly urged that a general vocabulary test be utilized to carry the burden of measuring the verbal factor when that is wanted.

Mathematical Tests

Mathematical and arithmetical computation are rarely mentioned among job-analysis traits in connection with training, but as will be seen in chapter 1, they are given place in the combat studies. Both mathematical and numerical-computation tests were found to be highly valid for the navigator, slightly for the bombardier, and slightly or not at all for the pilot. In this connection it should be said that the Arithmetic Reasoning test is not treated here nor in chapter 6 as a mathematical test, but rather as a reasoning test.

The only really unique feature of mathematical and numerical tests is their variance in the numerical factor. Although this factor is substantially weighted in all the mathematical tests, it is most unambiguously and satisfactorily measured by the Numerical Operations test, which also takes much less time. A factor identified as mathematical background is a substantial component of the Mathematics A (general mathematics) test, but it is slightly more efficiently measured by the Biographical Data Blank (navigator score). The conclusion is, that of all the mathematical tests tried, Numerical Operations has earned a permanent place wherever a measure of sheer numerical facility is wanted. No other mathematical tests, as such, seem to be fruitful in the selection and classification of air crew.

Reasoning Tests

Reasoning, as such, was never mentioned among the job-analysis categories. The concept of judgment was given considerable prominence, especially in connection with the pilot, however, and early attempts to analyze judgment from the psychologist's armchair led inevitably to an examination of reasoning tests of various types. While the Arithmetic Reasoning test was first developed to meet some of the needs of mathematical tests, it was shown later to be more of a reasoning test and so is treated in this group. It was realized early, however, that verbal and numerical tests were not valid for the pilot, and so nonverbal and nonnumerical reasoning tests were vigorously sought that might be valid and might to some extent take care of the judgment category (see ch. 7).

Most reasoning tests, but not all, prove to have variance in a factor identified only as general reasoning, which is best measured by the Arithmetic Reasoning, Forced Landings, Pattern Comprehension, and Spatial Reasoning tests. It is almost as highly involved in the navigator criterion as the numerical factor, but apparently not at all in either the pilot or bombardier criteria. Thus, so far as this type of reasoning test is concerned, no progress has been made toward covering the important problem of judgment for the pilot.

A second reasoning factor, perhaps identifiable as the ability to reason by analogy, is strongest in the Figure Analogies, Pattern Reasoning, and

703320—47—55

the Gottschaldt Figures tests. It may have some small validity for the pilot, but its validities for bombardier and navigator are unknown.

A third reasoning factor, strongest in the Spatial Reasoning and Decoding tests, seems to have a negative validity for pilots. Nothing can be said concerning its validity for other specialties.

In general, all reasoning tests are difficult to purify. The type of material in which they are given—words, numbers, figures in spatial arrangement—tends in itself to introduce an unwanted variance. Two or more reasoning factors are also likely to go together in tests. Fortunately, for navigator selection, both numerical and general-reasoning factors are highly valid. Fortunately, also, both are invalid for the pilot, so that a discriminating test is possible. For the bombardier, however, one factor is probably valid and one is not. The fact that we can supplement the Arithmetic Reasoning test by a pure numerical test, however, probably takes care of the bombardier and vocations similar in this respect.

Judgment Tests

The category of practical judgment, which was so highly regarded by instructors, received considerable attention (see ch. 8). It was demonstrated that verbally presented practical problems of the predicament type were factorially quite complex, including variances in verbal, general-reasoning, and mechanical-experience factors, as had been anticipated. But they also have variances in a new factor tentatively called planning and in a critical or judicial trait which may be called judgment. Early judgment tests owed much of their pilot validity to the mechanical component. This was natural in that the person with a good background of mechanical knowledge had an advantage both in solving the verbal-predicament problems as well as in pilot training. But subsequent findings have shown that the judgment factor, as such, has a positive contribution to make to prediction of the pilot criterion. One type of item that best measures this factor is of a common-sense-decision variety. Another is of the work-planning type. Tests of practical estimations of sizes, times, and distances are also related to this factor.

Near the end of the war, the hypothesis was being investigated that another component that might be called thought fluency is an independent contributor to practical judgment. The facile recall of pertinent experiences for possible use in everyday predicaments would give an individual more potential solutions from which to choose in a given unit of time. This hypothesis should be followed up if possible. The construction of tests of fluency using answer sheets presented new practical problems, but they did not appear to be insurmountable.

A judgment test of the verbal-problem type was in the classification battery during the last year of the war. Though the early tests of this kind were factorially complex, experience showed that they could be somewhat purified. The test is time-consuming (requiring an average of at least 1 minute per item) and reliability is typically low. There seems little likelihood that reliability can be substantially improved within practical time limits, but the contribution to the validity of the pilot stanine is probably undeniable, in spite of low reliability. The test should remain in the battery until something better covering the judgment factor is demonstrated.

It seems probable that other experiential background than the mechanical is important in connection with what the instructors call practical judgment. The general-information tests may owe a part of their pilot validity to this source, though analyses have shown no components in these tests except the mechanical experience and another factor identified as pilot interest. In this connection, as well as for other reasons in the study of practical judgment, it might pay to collect from the training fields instances of good and poor judgment with a view to seeing what kinds of information would have been pertinent to a successful solution to the practical problem. This would be another source of items for information tests.

Foresight and Planning Tests

This category, which was given prominent place in analyses of the pilot's job, seems to need drastic revision as a concept. The several tests that were developed to measure the hypothetical trait failed to show any single element that was common to them all, when they were statistically analyzed (see ch. 9). The a priori grouping of tests in this area under the categories of pathway planning, economical procedures, and planning by deduction also failed to find support in subsequent analyses. Tests in this area are generally complex factorially, with perhaps strongest affiliations with the general-reasoning and perceptual-speed factors, but some of them did, in fact, have in common a new factor which was called planning. It is not unique to planning tests, however, having been found in some judgment tests. The slightly promising validity of the planning factor for pilot training justifies further work toward better identification of it and improvement of tests for it. Its validities for bombardier or navigator are territories yet to be explored.

Integration Tests

Integration tests were developed with the intention of measuring by means of printed tests the most valid aspects of the very successful Complex Coordination test (see ch. 10). Ability to integrate activities in response to perceptually complicated situations was the working hypothesis. While the key to the valid intellectual component of the Complex Coordination test proved to lie elsewhere, integration tests were found to contribute three new factors apparently having to do with different aspects of mental sets—persistence, adaptability, and span. The first two promise some pilot validity, while the third seems to have a negative validity.

703320—47—56

Their validities for navigator and bombardier are unknown. From the superficial aspects of the tests one might expect some navigator validity.

Integration tests are relatively free from factors already well known, except for general reasoning in one or two, but some exhibit more than one of the integration factors. The relative uniqueness of these tests should appeal to those who are looking for selective tests where variances in mental-set abilities seem important. For the sake of determining factor validities alone, the tests best representing those factors should be validated for bombardier and navigator, and confirming studies of pilot validity are needed.

Memory Tests

A systematic exploration of the area of memory tests that superficially resembled memory tasks in aviation revealed three separate memory factors, all probably valid to a small degree for the pilot (see ch. 11). The first factor was called paired-associates memory, since it is characteristic of paired-associates memory tests. It is probably identifiable with Thurstone's rote-memory factor. The second is a visual-memory factor, which Carlson had previously discovered. It is characteristic of tests in which both learning and recognition tests are pictorial and identical. The third factor was confined to two tests in which object and name are paired. There is some indication of navigator validity for some of the tests in this group, but further study of this is needed.

One test, Memory for Tactical Plans, showed pilot validity strikingly beyond that attributable to any known factors, including the three memory factors mentioned above. Whether the valid component is a memory or a nonmemory factor is not known. If a memory factor, whether it is associated with the 2-hour delay between learning and recall quiz or to the type of material (verbal instructions) is unknown. It can possibly be identified with the integration I factor (persistence of mental set), or integration I may be a memory factor—memory for instructions. Such useful unknown variance is a challenge to discover its nature and to capitalize upon it to a greater extent. This test could be improved for the pilot by reduction of its verbal variance. Its variance in visualization is dispensable, also, since it is covered better in other tests.

Visualization Tests

"Visualization of flight course" is one of the 20 prominent job-analysis traits in the pilot list. The concept inspired very few test ideas, but results of research have shown abundantly that a factor of visualization does indeed exist, and that it is a significant component not only of the pilot criterion but also of the navigator and bombardier criteria as well (see ch. 12). Whether or not the concept of "visualization of flight course" is close to the meaning of the visualization factor is an unanswered question, but one that does not particularly matter unless it contains important aspects that the factor misses.

A variety of tests measures it, some of which, Mechanical Principles, for example, were unsuspectingly developed for quite different purposes. The ability seems to involve visual-thinking activity in which objects must be moved or transformed in order to solve a problem. It is distinct from visual memory on the one hand, and from spatial relations on the other. Its separation from visual memory and the space factor, which has traditionally been described as spatial visualization, is one of the important psychological findings of the program. It is not a feature of planning tests, as had been expected.

No visualization test yet developed is pure for the factor. General reasoning is the most common secondary factor, and some mechanical tests have visualization as their secondary factor. The general-reasoning variance might be eliminated by ridding the test of the more complicated and difficult items. It is believed that oral, verbal presentation of the items is probably best, with some control of the time factor for each item. Some of the later developed tests made use of these suggestions. It is important that they be analyzed to test the implied hypotheses.

Mechanical Tests

Many varieties of mechanical tests were tried out in the program—Mechanical Information, Mechanical Principles, Mechanical Movements, Mechanical Comprehension, Mechanical Functions, Tool Function, and Physical Principles, not to mention Pattern Assembly and Pattern Comprehension (see ch. 13). Both mechanical insight and mechanical experience were thus surveyed from many directions. Familiarity with machinery, a natural "knack" for understanding new mechanisms, or an inborn aptitude for success in mechanical tasks, if such there be, were all thought to be covered by the range of tests included.

Experience bore out the expectation that mechanical tests would be valid for air-crew selection. They were highly valid for the pilot and to a small but useful degree for the navigator. It was surprising to find them not valid for the bombardier, in view of his dependence upon understanding of the bomb sight and the autopilot in his training.

The secret of the validity that can be called unique to mechanical tests lies in the mechanical-experience factor. It would appear that any purely mechanical aptitude is largely an acquired trait, because the factor is by far the strongest in the Mechanical Information test. This fact does not deny that mechanical experience is gained somewhat in proportion to the power to gain it when equal opportunity is available. Opportunities for mechanical experience are not equal, however, and hence it would have been reasonable to expect two factors, one attributable to the power to learn and the other to opportunity. It may be, in spite of inequalities of opportunity, that two such factors are inextricably mingled in the analysis. At any rate, there is no common variance in mechanical tests to be accounted for, other than well-identified nonmechanical factors and the factor so characteristic of mechanical-information tests. Supporting the experience hypothesis in naming the factor are the very substantial loadings for it in the Biographical Data Blank (pilot score) and in the General Information test.

The development of mechanical tests for air-crew selection and classification may be regarded as having reached a satisfactory status. A mechanical-information test, which may also well include items on auto driving, tool function, and motor malfunctions, best represents this area. Not included in the survey of this area were tests of manipulations of tools and machines. These abilities call for psychomotor tests. It is believed, from the limited factor-analysis experiences with psychomotor tests which have been mentioned in this volume, that no manipulative abilities peculiar to tool handling or machine operating will be found that are unique. More general psychomotor abilities will probably take care of this aspect of mechanical endeavors.

General Information Tests

General-information tests were found to be valid to a practical degree for both pilot and navigator, but only for the pilot did they make a unique contribution (see ch. 14). Their most valid types of item for pilots were on aviation information, flying information, mechanical information, active-sports-and-hobbies information, and automobile-driving information. The unique contribution was in a factor called pilot interest. Two substantial secondary variances were in the verbal and mechanical-experience factors which are covered by other tests.

In addition to the coverage of pilot interest, general-information tests were developed in connection with other hypotheses, e. g., masculinity-femininity discrimination, and leadership qualities. With what effectiveness they will discriminate men on these variables is still to be determined. These aspects of information tests should be explored further.

Perceptual Speed Tests

The concept of perceptual speed, introduced by Thurstone, seems likely to endure, judging by the ease and the frequency with which the factor by that name is verified, and the ease of constructing pure tests for it. It appears to be a significant variable in all of the air-crew positions and it is well measured by three tests in the classification battery (see ch. 16).

Tests of a clerical type—such as Graph Reading, Table Reading, and Number Reading—were treated in this area because they probably have a substantial variance in the perceptual-speed factor and because they are highly speeded tests. Judging by their face appearance and their very high navigator validities, they probably have even greater loadings in the numerical and perhaps spatial-relations factors. In spite of their shortness, and consequent limited reliability, because of their high navigator validities and their simplicity of administration and scoring, they deserve serious attention whenever navigators must be selected in limited time. They are probably more complex than is desirable for general classification testing.

Form Perception Tests

Tests in this area were developed to help to round out an exploration of the general field of perceptual tests rather than to meet any recognized job-analysis requirement. Two pattern-formation tests had small to moderate validities for pilot selection, but this could be accounted for in terms of already known factors adequately covered by better tests (see ch. 17). Two completion tests had almost no pilot validity but probably contain new factors not already listed in this program. Pattern-analysis tests promise low to moderate pilot validity, the Gottschaldt test in particular. The latter is factorially complex and might be cultivated for its possible variance in the second reasoning factor. There is evidence, from outside the program, that geometric-illusions tests contribute a new factor. Neither factorial nor validity information is available concerning them from within the program. Should their unique contribution be found fairly strong and univocal, they should be validated against training criteria in several specialties.

Size and Distance Estimation Tests

In connection with these tests the most significant finding is the discovery of a length-estimation factor which is undoubtedly valid for the pilot and perhaps for other air crew (see ch. 18). After considerable effort to make size-estimate items realistic and to obtain judgments in aviation-material settings, it was found after all that the simplest form of item—namely, single straight lines—seems best. The factor was also found in other perceptual tests not designed as space-estimation tests, for example, the Pattern Assembly test, a modified Minnesota Form Board test.

Angular-judgment tests may be found to measure another space-judgment factor that is valid for selection. Tests of this type were developed late in the program and have not been analyzed or fully validated.

Problems of the generality of the ability to judge extents in one dimension with distance judgments in three dimensions, or of judgments of short lines seen close by with judgments of lines at great distances, have not as yet been solved. Discussions of theoretical considerations of these problems will be found in report No. 7 on motion-picture tests.

Spatial Tests

These tests (see ch. 19) proved to be very highly valid for the pilot, moderately so for the navigator, and significantly so for the bombardier. The discovery that a large part of the validity of psychomotor tests can be attributed to the spatial-relations factor and that this factor can be measured as well, and perhaps even better, by means of printed tests, such as Instrument Comprehension II and possibly Planning a Course, is of considerable practical importance.

The separation of space appreciation, as such, from visualization may be regarded as a distinct step forward in mental measurement. The finding of a second space factor which is slightly valid for pilots and not measured by any classification-battery test is also a positive gain. The presence and nature of a third possible space factor needs considerably more investigation before the factor can be accepted as real or its characteristics defined. At present, it seems to be strongest in two psychomotor tests, Two-Hand Coordination and Rotary Pursuit, but its substantial presence in two printed tests (plotting tests—rights score) suggests that like the first space factor, if it is genuine, it also is amenable to measurement by printed tests.

Orientation Tests

Job analysis for the pilot stresses ability to maintain orientation in space as an important perceptual trait. This inspired a large number of test ideas, and a great many kinds of orientation tests were developed (see ch. 20). Insufficient work has been done with them to determine whether or not they have any unique valid variance to offer. Analyses thus far have revealed the already well recognized factors of perceptual speed, spatial relations, and visualization in them. They deserve further study if only to identify their unknown nonchance variance and to determine its validity. Orientation tests have shown considerable pilot and navigator validity, but it is possible that the factors just mentioned and others that are known will account for most of that validity. There is still the possibility that a compass-orientation factor can be brought out, but its bearing upon air-crew selection is problematical. Many of the later developed tests were not yet analyzed or validated. This should be done.

Tests of Set and Attention

Tests of attention had quite low pilot validity but seem to have more promise for navigator validity (see ch. 21). Whether the latter can be attributed to an attention factor or factors is unknown. An attention factor in such tests as were used had been reported by earlier investigators.

The hypothesis of a change-of-set trait was found to be without support. Tests developed to measure it were almost entirely lacking in communality. This finding is in harmony with prewar results on perseveration tests, if we may regard change of set as the opposite of perseveration.

The integration tests, since they apparently turned out to be measures of various aspects of mental set, belong in the present category. At first thought, there is an apparent discrepancy between the finding of the factor integration II (adaptability of set) and the failure to find a change-of-set factor. Integration II was found strongest in the following Directions test. The difference in the two kinds of test must be that in the latter the examinee is led to expect changes in set, and he can or cannot adjust himself to them. In the change-of-set tests, every condition leads him not to expect a change of set. Changes that do occur are probably more or less fortuitous. Although the reliabilities of change-of-set tests could not well be determined, one might expect them to be rather low, except as they depend upon other factors. Because of the disparateness of the kinds of material and tasks in different change-of-set tests, other factors likewise failed to appear. It was to avoid the appearance of inconsistency that integration II was defined provisionally as a matter of adaptability of set rather than flexibility of set.

The job-analysis category of attention is called division of attention. No test was designed specifically for that aspect of attention, but it is possible that one or more of the integration factors come close to the concept of division of attention; but which ones they may be cannot be decided offhand. By definition, as at present envisaged, integration III would seem logically closest, but that is the one that appears to have some negative validity for the pilot, and this does not agree with the job-analysis idea. Further study of what is actually meant by division of attention of the pilot in action is needed. It may well be that apparatus tests are demanded for it.

Personality Inventories

A number of commercial personality inventories of the questionnaire type were tried in order to assess the general promise of this kind of approach to suitable temperament qualities in flying trainees and combat aviators (see ch. 23). These tests generally failed to exhibit pilot validity for the training criterion when scored with the already established keys. There were exceptions for three inventories, in which keys gave significant validities between 0.10 and 0.20.

The number of valid items yielded by these inventories was generally small, and sets of apparently valid items based on two independent samples did not contain many items in common. Cross-validation of empirical keys derived from sets of valid items generally showed failure of the validity of the items to hold up, with one or two exceptions. It is urged, however, in view of the paucity of valid, printed, temperament tests, that the most valid items (preferably those with correlation coefficients with the criterion that are significant at the 1 percent level) be selected from all inventories and assembled for future validation as a single test, against all training criteria. Validation against combat criteria, which would have been most desirable for all temperament tests, was not practicable, and since the close of the war is, of course, impossible.

Other experiences with tests, as well as these with inventories, lead to the conclusion that one stands the chance of most positive gains by designing tests and items for specific purposes. An exception to this generalization is that tests known to measure well-recognized factors are adaptable wherever those factors are significant components of the criteria one desires to predict. One might, therefore, be able to construct many new questions more suitable for air-crew selection, if the questionnaire type of item is what one wants. It is possible that better types of items would serve the same purpose; types that do not have some of the same objectionable features. Under the present circumstances, however, it seems desirable to follow out the suggestion of the preceding paragraph, to have the satisfaction of knowing whether questionnaires do or do not have anything unique to offer; and, if they have anything to offer at all, whether other types of items can serve the same purpose.

Preference Inventories

Vocational interest and other preference inventories were studied in the same manner as personality questionnaires with quite similar, though less promising, results (see ch. 23). It is probable that further work on this type of instrument would be unfruitful as compared with other types of interest tests developed specifically for the AAF with rather gratifying results.

Projective Tests

Neither of the best known projective tests—Rorschach and Thematic Apperception—showed promise of a practical degree of validity for selection to meet a pilot-training criterion (see ch. 24). Adaptation of the Rorschach to group testing yielded results that seem to be worth following up. It is believed, however, that new ink blots developed for the purpose would be desirable. Adaptations of the thematic principle to printed group tests were tried in a number of ways. None that was validated against the pilot-training criterion was promising. Validation against a more pertinent criterion in which psychoneurosis plays a more common role would be most desirable, for these and other temperament tests.

Observational Methods

These constitute a heterogeneous list of procedures in which the evaluations are in the form of ratings by observers. Observations of students were made under various kinds of situations; during the taking of psychomotor tests, during rest periods between psychomotor tests, and during interaction-test situations. Typically, a number of traits were looked for and rated (see ch. 24).

Like all personal ratings of traits, these evaluations suffer from the subjectivity factor of the human observer. Individual differences among raters both as to numerical values and as to qualities emphasized are bound to occur even under the best indoctrination. Wherever significant validities for such ratings may be found, it is desirable to set up hypotheses as to the significant traits involved and then to seek objective, quantitative indicators of the same traits.

An incidental finding that is of interest in connection with observations was that ratings of goodness of physical appearance of students as they took their group tests had zero validity (0.03) against the pilot-training criterion. The correlation of these ratings with evaluations based upon personal, quasi-psychiatric interviews was so low (0.13) as to indicate that physical appearance as such has a minimal bearing upon the interview result. The reliability of the interview rating is unknown.

Clinical Type Procedures

Recognized as being impractical in a mass-testing program, clinical-type procedures which emphasize a global approach to personality were nevertheless given an experimental trial (see ch. 24). Ratings of probable success in training were made, taking into account a large mass of data concerning each student. The sources of information were an hour interview, scores from Rorschach and Thematic Apperception, and similar tests, and observational data. The over-all ratings showed practically no validity for predicting the training criterion. This was true in spite of the fact that intimate observations of men in pilot training had led to the belief that personality traits were of considerable moment in students' adjustments to training, and probably to success in training.

While this finding does not also answer the question whether the psychoneurotic-prone individual can be detected by similar methods, the costliness of the procedures in personnel time and the subjectivity of the evaluations are good reasons for hesitation to pursue the matter further in a program in which an important goal is objectivity. Proponents of the global approach and the clinical-type procedures that go with them may be able to say that the approach was not given a complete trial, and members of the program will be ready to admit that this is so. Whenever any such project fails it can always be said that the optimal procedures were not employed. The reply to such an assertion is that relatively minor shortcomings should not be responsible for complete failure. There may be room for debate, of course, as to whether a shortcoming is minor or is a keystone.

Masculinity-Femininity Tests

Explorations were started with masculinity-femininity tests of the traditional type (Terman-Miles and Goodenough) and of information tests presumed to be discriminatory of masculine versus feminine characteristics (see ch. 25). The stimulus for this was the hypothesis, derived from observations in combat zones, that the masculine-type man is on the whole a better leader and also probably less psychoneurotic-prone. It was also believed that the trait would show up, though to a less degree, as between graduates and eliminees in training.

The traditional tests, when applied to aviation students, have shown very low internal consistencies. A single sex would naturally be expected to show less internal consistency than a mixed group, but there should still be significant differences within each sex. Several information tests were developed but none was validated against the training criteria. They should also be item-validated by carefully sampling large groups of males and females in representative populations in the age range of aviation students, in order to test the hypothesis more fully. The failure of the masculinity score of the Guilford-Martin Inventory of Factors GAMIN to correlate with the pilot-training criterion (see ch. 23) is one indication of the probable result with other masculinity-femininity tests.

Carefulness Tests

These tests were developed on the basis of a hypothesis arrived at after months of study of navigation training by the psychological research project (navigation); the hypothesis was that critical errors are made because the navigator is careless in minor details. Four tests of a complex clerical type, with some face validity for navigation, were developed and analyzed (see ch. 25). It was a striking finding that their error scores yielded a new common factor practically independent of the rights scores for the same tests. This seemed to be a carefulness factor, such as the hypothesis had called for. Its validity for the navigator has not yet been determined and is one of the urgent postwar navigator-selection problems. Incidentally, the finding of such divergent communalities between rights and wrongs scores for the same tests led to renewed scrutiny of similar phenomena in other tests and a reexamination of the general utility of formula scores.

Fear and Tension Tests

Tenseness and fear and apprehension are given as two separate categories in the job-analysis list of pilot traits. The theory behind test construction implicitly treats them as one, though tenseness as a symptom is measurable better by apparatus tests, while tendencies to fear are amenable also to testing with printed instruments.

Tests based upon expressions of opinions and attitudes, aimed at assessing generalized fear and nervousness, were only minimally satisfactory when validated against the pilot criterion (see ch. 25). The variance was probably unique, but the correlation with the pilot criterion so low that the contribution to the composite score was insufficient for practical gain. The exact source of even this low validity is yet to be determined. Judging from analogous results, we might find it not to be fear or nervousness at all. But whatever it is, the effort to define it might be worth while.

Confidence Tests

Lack of confidence is mentioned in the list of traits significant in navigator and bombardier training, but not among those for the pilot. Clinical observations by members of the program, however, tend to relate the trait also to pilot-training success.

Of several methods tried, only one, which was based upon self-ratings of performance on psychomotor tests, was valid for pilot training, and this to a practical degree (see ch. 25). The score was the divergence between the student's rating and the actual level of performance. The more realistic the judgment, the greater the student's chances of graduation. It is not known whether this validity should be attributed to self-confidence or to some other aspect of the rating. The subjective element in ratings is again a cause for restraint against using the device for selection and classification.

Social Intelligence and Leadership Tests

The need of measures to show how adept and how well adjusted an individual is in social relationships is urgent as one aspect of the problem of the selection of leaders. Internally consistent sets of items were developed for tests in this area, but the nature of the communality was not established, and no validation data are available (see ch. 25). Although the program was called upon to supply a "mental-alertness" score and a flight officer examination to be used toward the discrimination of officer material, not a great deal of research was directed toward the specific officer problem. In view of the general importance of the problem to the Air Corps, it would seem that much more attention should be given to it.

Motivation Measures

Motivation, including interests and attitudes, was mentioned as a significant trait in connection with all three assignments in training and in combat. The subject has several aspects which were treated in chapter 26, and several types of instruments were tried out to meet the apparent needs.

The student's own exprssion of his preferences for training and of his strength of interest in the types of training given him were used along with aptitude scores to determine classification, until the last year of the war. These expressions of interest failed to correlate with graduation-elimination to a practical degree except for the navigator. For the navigator, also, expressions of interest correlated substantially with valid navigator tests and with the navigator stanine. It is believed that the navigator's ratings of interest were, therefore, made in a more enlightened manner, with greater self-knowledge of potentialities. That this made the ratings themselves more valid is an interesting hypothesis.

Tests that relatively objectively assess a factor of pilot interest—the General Information (pilot score) and possibly the Biographical Data Blank (pilot score)—were valid to a practical degree. No factor of navigation interest, as such, emerged. Interests in academic work, more particularly in mathematics and numerical work, probably served a function in the selection of navigators comparable with that of pilot interest for the pilots. These interests may have been summarized, or incorporated in the mathematical background factor found in the Biographical Data Blank (navigator score) and in the Mathematics A test.

Interests and attitudes of fighter and bomber pilots proved to be sufficiently divergent and stable that it was possible to assess them by means of satisfaction tests.

Tests designed to assess combat readiness did not show significant validation against the training criterion. The internal consistency was generally low. If this supposed trait is to be measured, it is recommended that a new start be made.

Biographical Data Blanks

The outstanding success in deriving valid scoring keys for this type of information is proof enough of the utility of this kind of test (see ch. 27). Furthermore, it has unique contributions to make. The pilot validity for the pilot score is due in large part to the variance in mechanical experience, a component that could well be dispensed with in the blank. There is much in the pilot validity that is yet to be accounted for, and no hypothesis for it has as yet been accepted.

The question of how much, if any, validity of the scores is affected by guileful falsification has apparently been answered satisfactorily by means of an experimental study reported in chapter 27.

There is now little doubt of the need for this type of instrument in the classification battery. Scoring keys were in the process of construction for flight engineer, and it is likely that they could also be profitably constructed for radar observer and other specialists, including the bombardier. Possibilities of new items for pilot and navigator are real, as late item validations have shown. It is probably desirable to keep any such instrument up to date by periodic revalidation of items. Once the secrets of validity are unfolded, as in the finding that the pilot score depended much upon the mechanical factor for its validity, tests may be constructed to cover the same trait more objectively.

Other Job-Analysis Categories

The listing of tests above does not exhaust the job-analysis categories. The reasons will be clear as they are mentioned. Estimation of speed is not a trait amenable to printed testing; it requires apparatus presentation and control, and is best adapted to the motion-picture medium. The sense of sustentation obviously requires apparatus. Speed of decision and reaction was subjected to testing first by means of the Discrimination Reaction Time test. The analysis of that test shows that its validities for all three air-crew positions can be almost fully accounted for by known factors without resort to a decision factor or a speed factor. There is much nonchance variance in the test still unaccounted for, however, and a small part of its pilot validity. Thurstone's finding of a speed-of-decision factor (1) calls for further exploration of this area. The types of tests in which he found it, however, were very different from the Discrimination Reaction Time test. An effort was made to duplicate this test in printed form (see ch. 19), but whether it will carry all the factors present in the apparatus test remains to be seen.

The job-analysis traits mentioned under the heading of coordination and technique are suggestive of apparatus tests. No printed tests were even suggested to measure them: coordination, appropriateness of controls used, feel of controls, smoothness of control movement, and progress in developing technique.

Among the temperament categories, absence of tenseness has generally been regarded as an apparatus-test subject. By the direct observational methods involving ratings of tenseness during performance on psychomotor tests, little success was achieved. Ratings varied so much from observer to observer and from task to task as to justify little confidence in the validity of the ratings. Absence of confusion and nervousness has also been regarded as subject matter for apparatus tests. Direct observations seemed to give promise of validity in pilot selection, but objective scores on the task were even more promising and it is unknown whether the ratings added anything new. The two were highly correlated. The validity of both might represent variance in coordination or skill rather than in a temperament variable. The category of suitable temperament is indeed ambiguous. It can be interpreted broadly enough to encompass all other temperament categories. Until better defined, it is of no use as a guide to test ideas.

THE SELECTION OF BATTERY TESTS

It may be of value to record here some recommendations as to the compilation of future test batteries for whatever purpose they may be needed but more particularly for air-crew classification. There may well be differences of opinion on theory and procedure as well as upon specific recommendations. The ones made here are not by any means the only ones that could be made. They represent the result of the experiences of many investigators but the biases of a single interpreter.

Much discussion was held in the later months of the program concerning alternative batteries. Alternative batteries, or at least replaceable tests, are most desirable when retesting is called for. During the early months of the program, adherence to the rule of no retesting was strict. After excombat returnees began coming in large numbers and applying for air-crew training, retesting of them was made an important exception to the general policy. Another reason for alternative instruments is the matter of possible coaching, either by those who had been through the tests or by an outside agent who has somehow discovered the nature of the tests. Experience did not lead to great concern on this point. The battery was very long and quite varied in content. Many tests were not very adaptable to coaching and others went through revision of content from time to time. At any rate, the question of what tests are interchangeable, and the existence of a large number of experimental tests that have never been used in the classification battery both call attention to the need for decisions as to alternate tests and possibly as to alternate batteries. Having made these decisions in advance, a new battery could be put into use on shorter notice.

The recommendations to follow will not go beyond tests that have been factor analyzed. Again, the knowledge of factorial content and of validities of factors will be an important source of guidance. It seems obvious to the writer that the central principle to be followed in setting up any battery would be to note first the factors that have variance in the criteria to be predicted and their relative amounts, then to select the best tests, in terms of strength and purity in those factors. Noting the factors with positive validity for the pilot, in table 28.17 for example, one might then go to table 28.15 and under each factor find the best test of it for pilots, taking into consideration its other substantial loadings, if no pure test is available. Especially to be avoided, if possible, would be the inclusion of a test with a secondary or tertiary loading in a factor that has negative validity for the criterion in question.

In what follows, no attempt will be made to propose complete batteries. Instead, each factor will be considered in turn, and the applicability of the best tests of the factor to the selection of air crew will be mentioned. Lack of information concerning validities of many of the factors for bombardier and navigator will prevent a completely satisfactory coverage. In some instances tests will be recommended for navigator selection contingent upon the later finding that the factor is a component of the navigator criterion. Very few references will be made to the bombardier, since very little is known factorially regarding the bombardier criterion, nor is there much prospect that much will be known. Listed under each factor category will be the recommended tests and following them the specialties for which recommended, if the factor is valid. Following the list some qualifications may be given. Code numbers do not specify the exact form of the test, since different forms are usually factorially alike, and it is also presumed that new forms will be developed, with improvements where called for. The tests are mentioned in approximately the order of choice, though in many instances they are indistinguishable as to probable value.

CAREFULNESS TESTS

Plotting Test, CE452 (wrongs score), for navigator.

Complex Scale Reading, CE454 (wrongs score), for navigator.

Plotting Accuracy, CE453 (wrongs score), for navigator.

NOTE.—The third in the list is not quite as pure as the others. Its secondary numerical variance should make it more valid for the navigator.

Integration I Tests

Signal Interpretation, CI656, for navigator.

Combat Planes, CI655, for navigator.

Flight Formations, CI654, for pilot or navigator.

Note.—Signal interpretation contains too much general-reasoning and integration III variances to be a good test for pilots. Combat planes contains too much of the same two factors, plus some unwanted verbal variance.

Integration II Tests

Following Directions, CP402, for pilot or navigator.

Code Analysis, CI653, for navigator.

Note.—The choice for pilots is the lesser of two evils. Both tests contain some verbal variance, but Code Analysis also has numerical and general-reasoning components, and what is probably worse, a substantial amount of variance in integration III.

Integration III Tests

Note.—No test in this group is called for in pilot selection, owing to the possible negative pilot validity of this factor. Code Analysis, CI653, is probably best for navigator, if the factor has navigator validity. Planning a Course, CI406, would be a suitable second choice, since its other substantial loading is in spatial relations, which is known to be valid for the navigator. If the planning factor is found to have navigator validity, then Planning Air Maneuvers, CI408, would be desirable. If Reasoning III has navigator validity, then Figure Classification, CI213, would be suitable. One defect of integration tests is that all are complex. A good hypothesis is needed regarding the nature of integration III and new tests to maximize that type of variance.

Judgment Tests

Practical Judgment (nonmechanical), CI301, for pilot.

Commonsense Judgment,[3] AAFQE JR Part 3, for pilot or navigator.

Practical Estimations I, CI308, for pilot.

Note.—The first of the three, Form CI301BX1, had low verbal variance but substantial general-reasoning variance. The second has probably more verbal variance than is good for a pilot test. The third (A form) had zero verbal variance, and its mechanical-experience and planning components are valid for the pilot.

Length Estimation Tests

Shorter Line, CP606, for pilot or navigator.

Nearest Point, CP607, for pilot.

Note.—The first is practically pure, and is recommended even though other tests have a greater loading in the factor. The second has small amounts of visual-memory, spatial-relations, visualization, and verbal

[3] This test was described in connection with the factor analysis and judgment tests in chapter 8.

(negative loading) components, all of which would add to the pilot validity of the test. Probably better than either will be the Estimation of Length test, CP531, but its composition is as yet unknown.

Mathematical Background Tests

Biographical Data Blank (Navigator score), CE602, for pilot and navigator.

Note.—Mathematics A is not recommended because of its length and its strong numerical variance. It is significant that mathematical background apparently has validity for the pilot, whereas the numerical factor has not. Such a distinction could not have been forecast from job-analysis information, or from test results, probably, without a factorial analysis. The only disadvantage of the recommended test lies in its slight negative loading with pilot interest.

Mechanical Experience Tests

Mechanical Information, CI905, for pilot and navigator.

Tool Function, CI906, for pilot and navigator.

Note.—Tool Function had a slightly higher loading in this factor in one analysis than Mechanical Information had in an average of several analyses. This finding needs verification before much attention is paid to it. Mechanical Information is apparently more pure, though the secondary loading of perceptual speed in Tool Function, if real, would not detract from either pilot or navigator selection, and could probably be removed in later forms.

Paired Associates Memory

Memory for Plane Silhouettes, CI503, for pilot or navigator.

Memory for Ships, CI504, for pilot or navigator.

Note.—Both tests have secondary loadings in perceptual speed and spatial relations, both of which are valid for pilot and navigator. The second has less verbal variance and might, therefore, be better for pilots.

Visual Memory

Plane Formation, CP805, for pilot or navigator.

Note.—New forms of this test would undoubtedly be better. Map-memory tests are not recommended, because there is probably too much verbal component for the pilot, and more paired-associates memory and perceptual speed than should be tolerated.

Memory III

Plane Name Memory, CI506, for pilot or navigator.

Memory for Landmarks, CI510, for pilot or navigator.

Note.—Both tests would cover both memory I and memory III quite well. They are otherwise pure except for a small amount of perceptual speed. They should be rid of memory I if possible. This seems unlikely, unless a better hypothesis of the nature of memory III appears.

NUMERICAL TESTS

Numerical Operations, CI701, for bombardier and navigator.

NOTE.—There is no close competitor either in terms of strength or of purity. It is easy to find numerical variance combined with either general reasoning or with spatial relations in other tests. The combination with reasoning would be satisfactory for the navigator but superfluous for the bombardier.

PERCEPTUAL SPEED TESTS

Speed of Identification, CP610, for pilot, navigator, or bombardier.

Spatial Orientation I, CP501, for pilot, navigator, or bombardier.

PILOT INTEREST TESTS

General Information (Pilot score), CE505, for pilot.

PLANNING TESTS

Planning Air Maneuvers, CI408, for navigator.

Practical Estimations I, CI308, for pilot.

NOTE.—The first test would be very valid for the navigator if both the planning and integration III factors are valid. There is too much of both verbal and integration III in it for a good pilot test. The second test combines planning with judgment and mechanical experience, both components of the pilot criterion.

PSYCHOMOTOR COORDINATION

Rotary Pursuit, CM803, for pilot.

Complex Coordination, CM701, for pilot.

NOTE.—A better test of the psychomotor I factor could be produced.

PSYCHOMOTOR PRECISION TESTS

Discrimination Reaction Time, CP611, for bombardier or navigator.

NOTE.—No really good test of this apparently is available. Fortunately, this test's stronger loading in spatial relations contributes to validity for both assignments.

PSYCHOMOTOR SPEED TESTS

Log Book Accuracy (no code), for navigator.

Marking Accuracy (no code), for navigator or pilot.

NOTE.—The first test has a strong secondary loading in the numerical factor and so is not recommended for the pilot. The second has a secondary loading in perceptual speed and so would do for either assignment.

GENERAL REASONING

Mathematics B, CI206, for navigator.

Forced Landing, CI652, for navigator.

Pattern Comprehension, CP803, for navigator.

NOTE.—The first test combines about equal loadings in reasoning and numerical factors, but the factors are about equally strong in the crite-

703320—47—57

rion. The second test has a substantial secondary variance in integration II. If that factor should be found correlated zero or negative with navigation, the second test is out of the running. The third test carries small loadings in perceptual-speed, verbal, and visualization factors, all of which do no particular harm for the navigator.

Reasoning II

Gottschaldt Figures, QP901A, might be recommended for pilot or navigator, but not enough is known regarding other nonchance variance. It is probably a very complex test.

Note.—There is no good recommendation here. Figure analogies is equally strong for the factor, but is exceedingly complex, having factors with zero or negative relation to the pilot criterion and factors with unknown relation with the navigator criterion. A much better test could be developed. Something along the line of Pattern Reasoning of the judgment-and-reasoning battery is recommended as a starting point.

Reasoning III

Note.—Only two tests compete for this list, and both are too complex, with secondary loadings whose validities are either zero or negative for the pilot, and are largely unknown for the navigator.

Space I, Spatial Relations Tests

Instrument Comprehension II, CI616, for pilot.

Complex Coordination, CM701, for pilot.

Dial and Table Reading, CP621, 622, for navigator and bombardier.

Discrimination Reaction Time, CP611, for bombardier and navigator.

Note.—An abundance of good tests here makes it possible to discriminate among the specialties to which the test is best suited, since the factor is a component of all three criteria. Secondary variances in the first test are reasoning II, visualization, with a small verbal admixture. For the second test the chief secondary component is psychomotor coordination, which is of no use apparently for either bombardier or navigator. The third test presents a strong numerical component and some perceptual speed, both of which are not amiss for bombardier as well as navigator. The last test has the strongest secondary variance in psychomotor precision, for which the pilot has no apparent use.

Space II Tests

Hands, CP512, for pilot or navigator.

Note.—A rather pure test for the factor. It is believed that the new test developed in the program will be even better, that is, Position Orientation, CP526.

Space III Tests

Plotting Test, CE455 (Rights score), possibly for navigator.

Note.—This factor is of such uncertain status that it is probably premature to recommend any test for it. The factor's pilot validity as indicated by limited early results calls for no test of this factor for pilots.

Social Science Background

Note.—No test will be recommended, owing to the uncertain definition of this factor. Certainly, the two tests with communality in it, Geography and History, have stronger verbal variance than they have in this factor. The negative pilot validity estimated for the factor is of interest, but hardly calls for attempts to measure it with negative weights in the composite. On general principles, the negative weighting of any trait score means adverse selection for that component; although that component may be negatively related to piloting a plane, it might be positively related to success as a plane commander.

Verbal Tests

Vocabulary, CI604, for navigator and perhaps for bombardier.

Technical Vocabulary and Information (navigator score), CE505C, for navigator.

Note.—Both tests are practically pure for the factor. For face validity it might be better to use the second test in preference to the first. Reading Comprehension is much too complex to be recommended.

Visualization Tests

Mechanical Principles, CI903, for pilot or bombardier.

Spatial Visualization I, CI204, for navigator or bombardier.

Pattern Comprehension, CP803, for pilot and bombardier.

Mechanical Movements, CI904, for pilot.

Directional Plotting, CE455, for navigator.

Note.—A wealth of strong visualization tests exists—none, however, pure. The mechanical tests mentioned in this list give large portions of remaining nonchance variance to the mechanical factor which can do the tests no harm for pilot selection. The Spatial Visualization I test is recommended for navigator in spite of its complexity in factors whose validity for navigator are unknown. The test's validity for navigator is so high that not much risk seems to be taken. The Directional Plotting test is one to watch in connection with visualization, since both rights and wrongs scores are loaded with it. The rights score also carries some numerical and space III variance, the latter being still a question mark; the wrongs score carries some carefulness variance, which is presumed to be related to the navigator criterion, but that fact must still be established. These qualifications would probably justify the striking of this test from the list pending further data. It is believed that tests developed very late in the war will prove to be much purer for the visualization factor. Their analysis is an important step remaining to be done.

SOME GENERAL EXPERIENCES AND RECOMMENDATIONS

It would be without a sense of completeness to bring to a close this volume, leaving unmentioned some of the more general lessons learned and the implications which came from the rich experiences involved in the printed-test program. Many of the fruits of these experiences have been presented in chapters 2 and 3, in connection with the brief account of research plans and procedures. Others have been pointed out or implied in chapters on tests as the opportunity arose. In the paragraphs to follow, some of these points will be repeated for emphasis and by way of summary. Some new points will be brought out because they arise from viewing the results in perspective and in retrospect. Many of them are not new ideas; they have simply been emphasized by the pressure of repeated and massive experience. Others emerge from the systematic approach through factor theory and factorial results. It is believed that those who shared in the experiences all recognize the problems entailed in the points to be mentioned. In places there may be some divergencies of interpretation and recommendation. It must be remembered that the following account is mediated primarily through one observer.

Factorial Batteries

Experiences, as recorded in the preceding chapter, in particular, tend to strengthen the conviction that there are a limited number of stable reference variables in both tests and criteria in terms of which test batteries can best be adapted to vocational predictions. The ideal test battery, then, should provide variances in all the factors that are significantly weighted in pertinent job criteria.

The traditional method of extracting a single predictive score from a composite of several tests has been based upon the multiple-regression-equation principle, each part of the composite being optimally weighted by least-square fit in order to maximize the multiple correlation between composite and criterion. It is common knowledge that the lower the intercorrelation of parts, the greater the multiple correlation. Factors have been found to be either independent, i. e., uncorrelated, or to have small positive intercorrelations. Pure tests of factors would accordingly yield maximum independence among parts of a composite. There are other ways of arriving at tests with unique contributions; any two tests with zero intercorrelation coupled with validity of both for a criterion will satisfy this demand statistically. The fact that two tests correlate zero with one another, however, does not assure that they are unique with respect to a third test, nor even that they are factorially pure. Even a complete battery of nonoverlapping tests might be arrived at by trial and error. It is believed that the shortest route to this goal is through factor analysis and factor hypotheses. It is also maintained that the approach is decidedly more meaningful and that this reduces the amount of trial and error to a minimum.

The classification battery in use at the close of the war included 19 different tests which yielded 21 different scores. The best evidence available indicates that this battery covers 12 known factors, plus two additional ones not identified (the hypothesized kinesthetic-motor factor in the rudder control test and a completely unidentified factor in the Biographical Data Blank). The number of factors that have positive loadings in the pilot criterion appears to be about 20; 17 that appear in table 28.17 plus the 2 mentioned above and 1 unidentified in the experimental test Memory for Tactical Plans. The navigator and bombardier criteria are very incompletely known, but the present knowledge, as represented in table 28.14, would indicate that a classification battery for the three assignments requires at least four additional factors. The battery just referred to probably accounts for but 14 of the some 25 factors that should be covered, not to mention a number of temperament factors that have not even been brought to light.

The coverage of 25 distinct factors would require at least 25 tests or scores, even with pure tests. The reason why 21 scores account for only 14 factors is the great amount of replication of coverage of the same factors, in some instances, by as many as 4 or 5 different tests. It is true that some factors are paired in tests in a way that meets the specifications required by similar combinations in criteria. An example of this is the combination of numerical and general-reasoning factors in the Arithmetic Reasoning test, and another is the spatial-relations and psychomotor-coordination factors in Complex Coordination. The difficulty is that the relative weightings may not be optimal and that in some other criterion the two factors may not be similarly paired. For a general-purpose battery that goes beyond three assignments, the needs for test purity are even more apparent. There seems no escaping the fact that any vocational battery requires a large number of tests for complete coverage, even at a minimal practical level of prediction. There are other reasons for large numbers of tests in batteries as the following discussion will also show.

Maximizing Discriminations in Classification

Implied in the previous discussion is not only the goal of achieving maximal selection for each assignment but also that of making the maximal differential selection, that is, optimal classification. This problem has not received the analytical (in the mathematical-deductive sense) treatment that it deserves. It is probably true that when the first of these two goals is approached the second is also nearer fulfillment, particularly for those individuals who are qualified for one assignment and disqualified for others. Of those who are qualified for more than one assignment, however, there still remains the problem as to which of two assignments is the better. The lower the disqualification rate, the more serious this becomes.

Criteria of job success are, of course, not mutually independent. They possess factors in common also. Being themselves much more complex than most tests, their degree of independence might be expected to range somewhat lower than that for tests, even relatively complex tests. We have no direct evidence as to just how independent the criteria for bombardier, navigator, and pilot training are. There were insufficient numbers of individuals who attempted more than one of these types of training. Even if there had been large numbers, the conditions for estimating the correlation among criteria would have been rather short of satisfactory. Very rough estimates of these intercorrelations have been made from the known factor components given in table 28.14. The correlation between any pair of criteria would be given by the sum of the cross products of pairs of factor loadings. With this method as the basis, using the incomplete information in table 28.14, we find that no intercorrelation exceeds 0.20. The intercorrelations between composite aptitude scores for the various classification batteries were typically above 0.50. In only one instance was any one as low as the region of 0.25, and that was between the navigator and pilot stanines. One reason for the failure to achieve the proper degree of independence was the fact of impure tests. Another was that each composite was set up independently of the others. An analytical solution to the problem would have made possible the derivation of weights for any one composite taking into account the goal of maximal discrimination in prediction.

Even if pure tests had been utilized, however, the intercorrelations of the stanines would not have been a minimum, owing to the intercorrelation of error variances in tests. A test score cannot, unfortunately, be weighted in a composite without also weighting its error variance. Deviations due to this source are in the same direction in all composites in which the score is a part. The best solution seems to be to use independently derived measures of each factor in different composites. This may call for more extra work than is justifiable; the cost in extra testing time and material may be too great. It would call for two or more different tests or scores for each factor that is weighted in more than one composite. There was enough of such duplication in recent classification batteries to permit this procedure, and in one or two instances this type of discrimination was utilized. In a battery of relatively pure tests, however, either two or more tests of each factor would be called for in some instances, or the same test could be given in separately timed parts, each score being used independently. The Numerical Operations test is already scored twice (front and back of an answer sheet). The parts would yield scores of lower reliability than would the two combined, but it is believed that the goal of reliability is of less importance than that of factor coverage, a point that will be discussed in the paragraphs following. The writer is definitely inclined to a large and varied battery of short, even though less reliable, tests in preference to a restricted battery of longer and more reliable tests. In the same time interval allowed for testing, there is probably much more to be gained by the use of a more comprehensive battery, even if something must be sacrificed in terms of reliability of single tests.

Reliability and Factor Variance

So much importance has been given traditionally to the matter of test reliability that some comments are needed here in defense of the statements just made. Arguments of a nonanalytical sort will be offered to support those statements. A little common-sense reasoning, based upon some fundamental theory about the factorial composition of tests, should suffice for this purpose.

In factor theory, the total variance of a test score is conceived as being composed of a number of independent component variances, additively combined. In terms of an equation,

$$v_i^2 = a_{1i}^2 + a_{2i}^2 + a_{3i}^2 + \ldots\ldots + a_{ni}^2 + u_i^2 + e_i^2 \qquad (29.1)$$

in which v_i^2 is the total variance of of test I, a_{1i}^2, a_{2i}^2, etc., are variances of factors 1, 2, etc., in test I, u_i^2 = any unique variance there may be in the test, and e_i^2 = the error variance in the test. The reliability of the test, r_{ii}, is the sum of all the nonchance variances, in other words, the sum of all terms in the right-hand side of the equation except the last.

The other side of the picture is the contribution of each factor to the validity of a test. The fundamental equation for the validity of a test in terms of its common-factor loadings is stated on page 839, but is repeated here in modified form to apply to test I:

$$r_{ip} = a_{1i}p_1 + a_{2i}p_2 + a_{3i}p_3 + \ldots\ldots + a_{ni}p_n \qquad (29.2)$$

in which r_{ip} is the validity of test I for pilot training, a_{1i}, a_{2i}, etc. are loadings of factors 1, 2, etc., in test I, and p_1, p_2, etc., are loadings of factors 1, 2, etc., in the pilot criterion.

It is fairly obvious from equation (29.2) that of the values in equation (29.1) only the nonzero loadings a_{1i}, a_{2i}, etc., have any bearing upon the size of the validity coefficient. They do also contribute to the reliability of the test, as can be seen from equation (29.1). But there is considerable freedom for r_{ii} to vary independently of them.

Let us assume that in test I only loadings a_{2i}, a_{4i}, a_{6i}, and a_{7i}, are positive and that of these, only a_{4i} *and* a_{7i} are positive in the pilot criterion. The validity of test I depends therefore upon these two factors and their loadings in test and criterion. From the first equation, it can be seen that r_{ii} could change, either increasing or decreasing, without affecting the validity of the test, if that change is brought about by changes of variances a_{2i}^2, a_{6i}^2, or in u_i^2. The reliability could even shrink to the sum of $a_{4i}^2 + a_{7i}^2$ without affecting the validity of test I. Even relatively pure tests, if the single factor loadings squared, a_{ji}^2 does not equal r_{ii}, there is room for loss of reliability without loss of validity. It is not known what the typical change in factor variances is when there is a change in r_{ii}. It may usually mean a proportional change all along the line. But it is deemed possible by proper control to devise new forms of tests in which variances of common-factor components can be increased or decreased independently and at will. Experiences in the program lead to the belief that we are entering a stage in test construction where that refinement will not be uncommon.

In this connection it is pertinent to cite one or two dramatic instances in which unusually low reliability has been accompanied by substantial validity. One is the Path Length test, CP628B, whose reliability was estimated as 0.25 (alternate-forms) and whose pilot validity was 0.23. Another is the Biographical Data Blank, CE602D, navigator score whose reliability was estimated as 0.35 (split-half) and 0.49 (test retest) and whose navigator validity was 0.23. A 15-item judgment test (part III of the Air Corps Qualifying Examination, AC10A) had a reliability of 0.36 (odd-even) and a pilot validity of 0.36.

From these cited results it looks as though the common advice that if a test shows validity one can forget about its reliability might be sound; at least in some tests. To see whether this type of case might be more general, a correlation was computed between the best estimates of reliability and of pilot validity for 74 tests. No test was included unless an alternate-form type of reliability was available or an odd-even type in a clearly power test. The coefficient was almost exactly zero. There may be constant errors, such as certain types of test, like vocabulary and mathematics tests, which have low validity but high reliability, that load the situation. A better controlled test of the matter could be made, but the fact is cited for what it is worth. Examination of lists of reliabilities of intellectual and perceptual tests described in this volume will show that they range from about zero to 0.97, with a median of about 0.80, the distribution being markedly skewed. Were we to hold out for a high minimum reliability level, many a useful test would not be utilized.

The question as to whether the variance components of a test that are not valid for a criterion should be in the nature of error variance or of variance in other, but nonvalid, factors, is an open one. We may assume for the sake of argument here that there is a real choice in the matter; that with the valid factor variance held constant, the reliability may equal that variance or it may be substantially greater by the addition of other common-factor variance none of which contributes to validity. In the first alternative, all nonvalid variance of a test is given over to error variance; and in the second, the nonvalid variance is divided between irrelevant but nonerror variance and error variance. What will be the different effects upon the contribution of this test to a composite under these two conditions? There are probably several effects, but one of them is that the test with secondary and perhaps tertiary factor variances will to that extent be selective in the directions of those additional factors whether we want that kind of selection injected into the composite or not. If all the nonvalid variance were error variance, no change of direction of selection would be entailed. From this line of thinking, therefore, it would seem preferable to use a less reliable, pure test to using a more reliable, more complex test, when both have equal projections on the valid factor that we wish to measure.

Validities of Pure Tests

In the search for valid pure tests, one finding is disturbing to the investigator who, following the traditional teachings on test construction, works toward maximizing the validity of each test. If the latter is the sole objective, we almost always end up with complex tests. An exception to this is when a criterion has a high saturation in some factor or factors, as the navigator criterion has in the numerical and general-reasoning factors. It turns out, as it did in the present program, that the most valid tests are complex; Reading Comprehension, Arithmetic Reasoning, Mechanical Principles, Biographical Data Blank (pilot score), Dial and Table Reading, and the Complex Coordinator test. By comparison, pure tests like Vocabulary, Speed of Identification, Numerical Operations, and Mechanical Information have suffered. All have been taken out of the classification battery giving way to other tests at one time or another because of lower validity coefficients. In the absence of thorough knowledge of intercorrelations or of factor loadings and their validities, the temptation is strong to do that very thing. In the light of such knowledge, all of these four, except the vocabulary test, were returned to the battery.

As another aspect to this matter, the attempt to improve a test by making item-validity studies also works toward complex tests. Any item may correlate with the criterion for as many reasons as there are factors in common between them. One item is valid because of factor A, another because of factor B, and still a third because of factor C. Or, as with total test scores, an item that is itself factorially complex has a greater likelihood of exhibiting significant validity and so of being retained. While the validity of the total score is thus raised, the uniqueness of the test is not thereby promoted.

A Technique for Test Purification

Tests have been improved with respect to validity by the procedure of item validation just mentioned. They have been improved with respect to internal consistency and hence reliability by item correlation against total scores. They should also be subject to purification by a similar process of item selection based upon item correlation with criteria of known factors. The selection can be both positive and negative; that is, acceptance of items that correlate acceptably with the factor to be maximized in the test, and the rejection of items that correlate to an unacceptably high degree with other factors. An arithmetic-reasoning test might be made more of a reasoning test and less of a numerical test by rejecting items that correlate too strongly with a Numerical Operations score. Other tests may be reduced in verbal variance by correlating items with a vocabulary score. Reading Comprehension might be rid of mechanical-experience variance by correlating items with a Mechanical Information score.

This technique was given a trial in connection with the test Mechanical Principles, in an effort to segregate the mechanical and visualization items, with no great success. The difficulty probably lay in the fact that the items had already been put through tests of internal consistency and some of them through an item validation, which as was said before works toward complexity. To be most effective, the purification by item correlation must begin early. It would be wise, in constructing the items originally, to have hypotheses about what introduces variance of one kind or another into an item and to strive for the kind of purity one wants. Awareness of what the factors are and in what kind of items they are likely to appear is a great help in this. The success that was achieved in keeping verbal variance out of a number of tests where it might well have crept in, e. g., Mechanical Information, some of the integration tests, etc., is an indication of what can be done.

Homogeneous Versus Heterogeneous Tests

During the course of test development in the program, it became apparent to many of the personnel, as it had been known from the beginning by others, that tests fall into two categories as determined by the homogeneity of their content. Homogeneity may be of different kinds. Homogeneity of form and content can be noted by superficial examination. Homogeneous items from this point of view will look alike in certain respects. Homogeneity of function is detected by means of item intercorrelation or of item correlation with some common criterion, such as total score in the test of which they are a part.

The similarity of form refers to the technical nature of the item, whether it is a multiple-choice, matching, or true-false, etc., type. Content similarity, by superficial inspection, refers to the material used—words, diagrams, forms, machines, colors, and the like. Each test is usually consistent in form and in content throughout, and the choice of either is determined by such considerations as conveniences of group administration, use of answer sheet, machine scoring, etc., and also by the nature of the task which it is believed will best bring out individual differences in the trait being measured. While these technical uniformities of items have some bearing upon the factorial composition of item variances and, therefore, upon the functional homogeneity of a test, there is much latitude for diverse functional nature within the same set of similar items. It is the functional homogeneity of items in which we are most interested here. The degree of functional homogeneity is indicated by item intercorrelations, but not as simply as one might think, as will be shown.

Item analysis in which the criterion of homogeneity of an item with other items in the test is the correlation of items with total provisional score on the test, tends undoubtedly toward increased homogeneity when the provisional score is itself relatively unambiguous, that is, factorially pure. When the provisional total score is itself factorially complex, it can be seen that the selection of items having greatest item-total correlations will not necessarily increase homogeneity in the factorial sense. Item validation, in which the basis of item selection is correlation of item with an outside practical criterion, is also likely to lead toward greater factorial heterogeneity, as was pointed out before.

The moral of this discussion is that if one desires valid, unique tests, one should proceed slowly in the use of the item-total correlation until one has a fairly unambiguous total score. The correlation of items with a job criterion may be used very well as an exploratory, preliminary step. The valid items should be scrutinized in order to derive a hypothesis as to their validity. Items that seem to fit the hypothesis in common may then be used as a cluster for deriving a provisional total score to be used as a new criterion for new item correlations, of items within the cluster as well as of newly constructed items which also appear to fit the hypothesis. In developing the General Information test, pilot score, for example, having found that the factor of pilot interest is the unique variance that the test has to offer, and that mechanical experience is another strong variance but well covered by another test, one should, according to this line of reasoning, reject items that correlate highly with the Mechanical Information total score (better yet, transfer those items to that test), then make an internal-consistency analysis of remaining items with a total score based upon remaining items as the criterion. If there are then some items with low consistency but still with significant pilot validity, effort should be made to understand the nature of any other valid factor that may be represented.

Apart from the goal of highly homogeneous, unique tests, if one has a test that is obviously heterogeneous and merely desires to maximize its validity, the route would not be through increasing internal consistency, but quite the opposite. Applying the common multiple-regression principles, one would strive to maximize the correlation of each item with the job criterion and to minimize the intercorrelations of the items. The selection should favor items of the type that will bring that about. This procedure, however, would seem merely to result in an extension of our ignorance to new valid territory, rather than to increase our knowledge of why tests are valid and therefore to improve our control over validity already achieved.

Power Tests Versus Speed Tests

Not a great deal was learned concerning the relative merits of power tests and speed tests, nor what effect working time as a determining factor has upon test results. The problem arose many times, but each particular instance of it was met as it occurred, without the benefit of any new general principles having been discovered or any general studies being made upon them. Some tests were administered with different time limits for the same material, and routine reliability and validity studies were carried through, but without results justifying any statement of generalizations. Only one or two rather disjointed comments, therefore, can be offered on this question.

In tests in which it was desired that all examinees attempt or respond to every item, the device of "pacing" the group was found useful. Reference to some test descriptions in earlier chapters will show that one or more times during the work on a test the administrator would break in with a statement to the effect that "at this time you should be working on item number X." In spite of this, there would still be a limited number who might not complete a test in the allotted time, liberal though it was. There is the other problem of the many who complete a test much earlier than the given limit and who have nothing to occupy them until the next test is called for. To meet this situation, it is recommended that many power tests include more items than are scored. The items at the end, beyond the last one scored, merely provide busy work for the rapid worker. The terminal, busy-work items may be those of low internal consistency and of high level of difficulty. This procedure would probably be most applicable to vocabulary and information tests, though it would work with others. One could then score only items that have been attempted by everybody, or down the list to a point where any desired percentage has attempted the last item.

If it is not known beforehand on the basis of soundest theory or empirical evidence whether a test is better as a speed test or as a power test, the recommended procedure would enable us to settle the point. If all individuals attempt items through the nth one, it would be desirable to determine the validity of scores derived from n items, n + 5 items, n + 10, n + 15, and so on as far as one desires to carry the study. Allowances for relation of validity to test length would need to be made.

The time problem deserves study from another aspect. It may well be that the validity of many a test is below its maximum because examinees are themselves too much in control of their working time. Pacing of the type mentioned may help to overcome this to some extent. Printing short sections of tests on each page and timing the test by pages is even better. Even more precise control of working time per item may be desirable in some tests. This suggests either tachistoscopic or motion-picture presentation of items in which the most stringent control can be attained. Empirical studies of this problem are needed.

Rights Versus Wrongs Scores

The reader who has followed the discussion of even a small number of the tests in this volume will almost certainly have noted the attention that has been given to rights and wrongs scores, apart from their mention in scoring formulas. Experience has repeatedly called attention to the importance of this. In a surprisingly large number of tests the dispersion of wrongs scores is relatively large, offering a basis for measurement of individual differences. There is also sufficient freedom in many tests for correlations between rights and wrongs scores to depart radically from −1.00. There is also the striking discovery that rights and wrongs scores may be functionally very different, giving indications of individual differences on quite different continua of human personality.

The implication of this is that in the development of any new test, consideration should first be given to the amount of correlation between rights and wrongs scores. If this is sufficiently different from −1.00, let us say −0.80 or higher on the scale (when corrected for attenuation), from then on they should be investigated as two test variables. They should be validated separately, have separate determinations of reliabilities, and each should be treated in factorial studies. In factor analysis, it would be best not to include both in the same matrix, if both are derived from the same set of items. A rights score from one half the test or one form and wrongs from the other half or form would be suitable, avoiding the emergence of a doublet factor unique to the two. When these procedures are carried out, it may be found that both scores are called for in using a battery, and that either or both should be weighted in one or more composite scores. If both should be called for in the same composite, it is recommended, again, that they be derived from independent sets of items.

Scoring Formulas

Closely related to the problem of rights and wrongs scores is that of scoring formulas. Experience shows that the automatic and indiscriminate use of correction-for-guessing formulas is to be severely condemned. While this procedure may satisfy the logic about probabilities of chance success with an item by guessing (and it is often doubtful whether this logic really applies to the case for which it was intended, as was pointed out in ch. 3), it may at the same time have quite serious effects that were not suspected and which, if known, would not be desired.

It might seem that the solution lies in deriving optimal scoring formulas, giving the rights unitary weight and the wrongs a weight *a*, which will maximize the multiple correlation between rights and wrongs, additively combined, and the job criterion. When this has been done, the a priori formula is verified less often than it is not, but validities of tests are raised very little by change in weight *a*. For 10 different tests in 1 study, the highest gain in validity for optimal weights was 0.04, and in most of them the gains were too low to be of practical value. As compared with validities for rights only, however, the optimal weights provided increases from 0.03 to 0.06. In two tests, positive weights, one as high as +0.479, were apparently best for the wrongs, with increased validities of 0.02 and 0.03 over those derived with negative, a priori, weights for the wrongs.

The use of the optimal-weight scoring formula leads to the conclusion that very large samples are required for the estimation of stable weights, and that the formula is extremely sensitive in some instances, giving very large changes in *a* for small changes in correlation coefficients. The wary investigator will be on the lookout for absurd results with the formula at times.

Cross Validations of Composites

Another, more general, caution in connection with multiple-regression weights is well worth mentioning. It has been recognized that a coefficient of multiple correlation as ordinarily computed represents the maximum amount of association between a pool of variables optimally weighted and a single criterion, and that this correlation is subject to some shrinkage when the weights are applied to predicting the criterion in a new sample from which the weights were not derived. The extent of this shrinkage is sometimes estimated by means of shrinkage formulas. These formulae are expected to indicate the amount of regression effect to be expected. Shrinkage formulas were rarely employed in the program, owing to lack of full confidence in them. Instead, as will be noted in chapter 24 and following, a cross-validation procedure was invoked to determine empirically whether a scoring key would maintain its validity when applied to a new sample. The results obtained by this procedure well justified the decision to expend the necessary effort. The amount of shrinkage is often surprising, but such a finding leaves one with the satisfaction of assurance that he knows the worst. The experience also leads to the suspicion that many a prediction composite may unwittingly rest on a shaky foundation. Aside from cross validation, another procedure that can be used is to test regression weights for statistical significance or to compare weights derived in two halves of a sample. Having gone this far, however, the cross validation entails little extra effort.

SOME GENERAL IMPLICATIONS

The preceding pages have presented a few of the many suggestions that emerge from experiences encountered in developing tests. They tend toward the more technical and statistical type of problem of limited scope. What can be said concerning the larger vistas of human mental measurement that surely must have been glimpsed from time to time? There have been moments when the vision has been less myopic than the detailed accounts of this volume suggest. It is hoped that the mere recital of facts about test after test, area after area, has also provided the reader with opportunities to share in the outlook that first-hand experiences have offered.

Those who have been close to practical problems of vocational selection or vocational guidance, well know that the development of useful techniques has been painfully slow and disheartening, and that the final limitations of effective prediction may have seemed to be very great. The ceiling of maximal accuracy of prediction may have seemed to be quite low. If this is so, it is believed that the measure of success achieved by the program which this and similar volumes represent, should materially alter the outlook for vocational-adjustment service for the better. There were probably very few at the beginning of the program who would have placed much hope in the prospect of selecting pilot trainees by means of printed tests alone with a degree of accuracy represented by a correlation as high as 0.50. That much accuracy has actually been accomplished by means of a battery including a total of only 150 items, as will be told in more detail in another Report. Nor would one have expected that by means of a longer battery which includes apparatus tests a validity between 0.60 and 0.70 could be attained for pilot selection, and an even higher validity for navigator selection. What has been done for these two occupations can be done for others, and, as may be seen in the latter part of chapter 28, the ceilings for pilot and navigator have not by any means been reached. The whole area of temperament is very much still open territory in this connection.

This degree of success has not been accomplished in four ordinary years, or with ordinary allotments of personnel, or with ordinary facilities and subjects. Only the crisis of a world war, unfortunately, could permit it to come to pass. Advances in other vocational areas will need somewhat similar concentration of efforts. Fortunately, much that has been learned in the Army Air Forces will readily apply elsewhere. Much of it, to be sure, will not apply without further research; but the groundwork has been laid. Much more needs to be done in fundamental, not immediately practical, research. The great richness of human talent and temperament has been emphasized as never before. The limitations of the IQ and the PQ have been thrown into bold relief. A society that wants a useful and dependable vocational assignment of its personnel must be ready to support the research that is required to satisfy that desire.

The volume should not be closed without pointing out the fact that the program was a highly cooperative affair. It may be regarded as an example of cooperative research and of what can be accomplished when trained individuals with common purposes and efforts tackle technical problems in the democratic way. There is probably no single contribution of which it strictly can be said that it is exclusively the brainchild of Captain X or it is the test constructed by Sergeant Q. Ideas were submitted to group discussion; test items to critics; plans were laid in conference; decisions were reached by agreement; specialized technicians each had a hand in the finished product. Things did not always progress as smoothly as this account may imply, but the essential idea was development of tests and projects as socialized ventures and the constant interplay of criticism and rebuttal. In this manner many a potential mistake was undoubtedly caught early, and the plan was the richer because of the multiple contribution. Because of the high proportion of creative work in research as such, and because psychological research, in particular, profits relatively more by a pooling of individual impressions, the cooperative approach has much to offer.

Having made these remarks about cooperation in general, it should also be said that all credit and the gratitude of those who have benefited or will benefit by the results of their efforts are due to the many individuals who loyally and unsparingly devoted themselves to one of the greatest adventures in human engineering.

BIBLIOGRAPHY

(1) Thurstone, L. L. A Factorial Study of Perception, Chicago: University of Chicago Press, 1944.

APPENDIX A

Mathematical Rationale for Chapter 3

In the derivations to follow, a number of simplifying assumptions are made which the reader may consider inappropriate. It is true that a condition such as equal item variances will never be met in practice. The chief usefulness of these formulas, however, is to give the test constructor a rapid way to predict statistics on some level other than the one on which he is working. The error introduced is small compared with the computational convenience of the simplified formulas. Compare, in this respect, the Kuder-Richardson formulae 20 and 21.

DEFINITION OF SYMBOLS

x_i = a deviation score in item i. It may take on various meanings: x_1, x_2, x_3 * * * x_n, respectively.

n = the number of items in the test.

σ_i = the standard deviation for item i. $\sigma_i = \sqrt{p_i q_i}$.

p_i = the proportion of the individuals who answer item i correctly. Correct answers are given a score of +1 and wrong answers or omissions a score of zero.

$q_i = 1 - p_i$.

σ_t = the standard deviation of the total test score.

r_{ij} = a product-moment correlation (phi coefficient) between any two items, i and j.

r_{it} = a product-moment correlation (point biserial) between item i and the total test score.

r_{tt} = a reliability (internal-consistency) coefficient for the total test score.

r_{tc} = the correlation between the total test score and an outside criterion.

r_{ic} = a point-biserial correlation between item i and an outside criterion.

$\bar{r}$ = an average (mean) of a number of coefficients of correlation.

DERIVATION OF EQUATIONS

Part I.—Item Intercorrelations

The correlation between an item and the total score, including the item, is given by the equation $r_{it} = r_i(x_1 + x_2 + \ldots\ldots + x_i + \ldots\ldots + x_n) =$

$$\frac{\Sigma x_i^2 + \Sigma x_i x_1 + \Sigma x_i x_2 + \ldots\ldots \Sigma x_i x_n}{N\sigma_i\sigma_t} \quad (1)$$

Dividing through by $N\sigma_t$,

(2)
$$r_{it} = \frac{\sigma_i + r_{i1}\sigma_1 + r_{i2}\sigma_2 + \ldots\ldots + r_{in}\sigma_n}{\sigma_t}$$

Assuming equal variances, i.e., constant difficulty for all items,

(3)
$$r_{it} = \frac{\sigma_i + \sigma_i \sum_{i=1}^{n-1} r_{ij}}{\sigma_t} = \frac{\sqrt{p_iq_i} + \sqrt{p_iq_i} \sum_{i=1}^{n-1} r_{ij}}{\sigma_t}$$

Rearranging and dividing through by $(n-1)$ we have the mean correlation between item i and all the remaining items in terms of the correlation between item i and the total, the standard deviation of the sum, and the mean standard deviation of the items:

(4)
$$\frac{\sum_{i=1}^{n-1} r_{ij}}{n-1} = \frac{r_{it}\sigma_t - \sqrt{p_iq_i}}{(n-1)\sqrt{p_iq_i}}$$

Summating for all items and dividing by the number of items,

(5)
$$\frac{\sum_{i=1}^{n} \sum_{j=1}^{n-1} r_{ij}}{n(n-1)} = \frac{\sigma_t \sum_{i=1}^{n} r_{it} - n\sqrt{p_iq_i}}{n(n-1)\sqrt{p_iq_i}} = \frac{\sigma_t\bar{r}_{it} - \sqrt{p_iq_i}}{(n-1)\sqrt{p_iq_i}}$$

Formula (5) gives us the relationship between the mean product-moment correlation between items (phi coefficient) and the mean product-moment correlation between item and total score including the item (point biserial), the standard deviation of the total score, and the mean item standard deviation.

The variance of a sum is given by the equation:

(6)
$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 + \ldots\ldots + \sigma_n^2 + 2r_{12}\sigma_1\sigma_2 + \ldots\ldots + 2r_{(n-1)\,n}\sigma_{(n-1)}\,\sigma_n$$

Assuming equal item variances, or constant item difficulty,

(7)
$$\sigma_t^2 = n\sigma_i^2 + \sigma_i^2 \sum_{i=1}^{n} \sum_{j=1}^{n-1} r_{ij} = np_iq_i + p_iq_i \sum_{i=1}^{n} \sum_{j=1}^{n-1} r_{ij}$$

Substituting an expression for $\sum_{i=1}^{n} \sum_{j=1}^{n-1} r_{ij}$ derived from (5), we have

(8)
$$\sigma_t^2 = np_iq_i + p_iq_i \left(\frac{\sigma_t\bar{r}_{it} - \sqrt{p_iq_i}}{(n-1)\sqrt{p_iq_i}}\right)(n-1)n$$

Canceling and assembling terms,

(9)
$$\sigma_t^2 = n\sqrt{p_iq_i}\,\sigma_t\bar{r}_{it}$$

Dividing both sides by σ_t,

(10)
$$\sigma_t = n\sqrt{p_iq_i}\,\bar{r}_{it}$$

Substituting the expression in (10) for σ_t in equation (5) and simplifying, we obtain the desired relationship between the mean interitem correlation (phi coefficient) and the mean correlation between items and total score (point biserial):

(11)
$$\bar{r}_{ij} = \frac{n\sqrt{p_iq_i}\,\bar{r}^2_{it} - \sqrt{p_iq_i}}{(n-1)\sqrt{p_iq_i}} = \frac{n\bar{r}^2_{it} - 1}{n-1}$$

Part II.—Internal Consistency of Total Scores

Substituting the expression in (10) for σ_t in the Kuder-Richardson formula number 21,

(12) $$r_{tt} = \frac{n}{n-1} \cdot \frac{n^2 p_i q_i \bar{r}^2_{it} - n p_i q_i}{n^2 p_i q_i \bar{r}^2_{it}}$$

Canceling and multiplying,

(13) $$r_{tt} = \frac{n\bar{r}^2_{it} - 1}{n\bar{r}^2_{it} - \bar{r}^2_{it}} = \frac{n}{n-1} - \frac{1}{\bar{r}^2_{it}(n-1)}$$

The internal-consistency coefficient of the test as a whole is thus stated in terms of the mean product-moment correlation between items and total score (point biserial).

Part III.—Spurious Item-Test Correlation

If r_{tt} equals zero, then

(14) $$\frac{n}{n-1} = \frac{1}{\bar{r}^2_{it}(n-1)}$$

Rearranging, and canceling,

(15) $$\bar{r}^2_{it} = \frac{1}{n}$$

Taking the square root,

(16) $$\bar{r}_{it} = \frac{1}{\sqrt{n}}$$

In other words, the mean product-moment correlation between item and total score including the item (point biserial) when inter-item correlations are zero is inversely proportional to the square root of the number of items in the test.

Part IV.—Validity of Items and of Total Score

The correlation between an outside criterion and the total test score given in terms of item scores, reads: $r_{tc} = r\,(x_1 + x_2 + \ldots\ldots + x_n)c =$

(17) $$\frac{\Sigma x_1 c + \Sigma x_2 c + \ldots\ldots + \Sigma x_i c + \ldots\ldots + \Sigma x_n c}{N\sigma_t \sigma_c}$$

Dividing the last term through by $N\sigma_c$,

(18) $$r_{tc} = \frac{r_{1c}\sigma_1 + r_{2c}\sigma_2 + \ldots\ldots + r_{ic}\sigma_i + \ldots\ldots + r_{nc}\sigma_n}{\sigma_t}$$

Assuming equal variances, or equal item difficulties,

(19) $$r_{tc} = \frac{\sqrt{p_i q_i}\ \sum_{i=1}^{n} r_{ic}}{\sigma_t}$$

Substituting for σ_t the expression in (10),

(20) $$r_{tc} = \frac{\sqrt{p_i q_i}\ n\bar{r}_{ic}}{n\sqrt{p_i q_i}\ \bar{r}_{it}}$$

703320—47—58

The validity of the total test score is thus stated in terms of the individual item validities and the item point biserials with total score.

Canceling terms,

$$r_{tc} = \frac{\bar{r}_{ic}}{\bar{r}_{it}} \tag{21}$$

In other words, when items are of the same level of difficulty, or nearly so, the validity of the total score is the ratio of the average (mean) item validity to the average (mean) correlation between item and total score.[1]

---

[1] This appendix was written by Capt. Lloyd G. Humphreys.

APPENDIX B

Factor Loadings, Communalities, Reliabilities, and Validities for Printed Tests Grouped Alphabetically

APPENDIX B—*Factor loadings, communalities, reliabilities,*

Test and Code No.	N_t	Ca	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Aerial Photographs, QP901A-IV	392	....	....	....	....	....	06	....	....	....	15	....	27	39	....
Biographical Data, CE602D (Navigator)	3000	....	....	....	....	....	....	42	−04	....	....	....	19	10	−14
Biographical Data, CE602D (Pilot)	3000	....	....	....	....	....	....	29	50	....	....	....	−20	14	−09
Block Counting, CP512A	392	....	....	....	....	....	04	....	....	....	36	....	−13	43	....
Code Analysis, CI653AX2	260	....	03	40	42	....	00	....	−07	....	....	....	29	06	....
Combat Planes, CI655AX5	266	....	57	12	28	....	10	....	−03	....	....	....	22	02	....
Competitive Planning, CI409AX2	372	....	....	....	33	36	....	....	−02	....	....	....	15	05	....
Complex Scale Reading, CE454A (Total Rights)	354	05	....	....	....	....	....	....	....	....	....	....	52	....	....
Complex Scale Reading, CE454A (Total Wrongs)	354	57	....	....	....	....	....	....	....	....	....	....	09	....	....
Cubes, CP512A	658	....	06	00	03	....	−13	....	−10	....	−03	....	18	31	....
Decoding, CI214AX1	202	....	....	....	00	....	....	....	....	....	....	....	12	36	....
Decoding, CI214AX2	1000	....	....	....	....	....	....	−06	−17	....	....	....	....	31	....
Dial Reading, CP622A	392	....	....	....	....	....	....	....	....	....	08	....	50	27	....
Dial & Table Reading, CP621A, CP622A	6000	....	....	....	....	....	....	15	−02	....	....	....	53	31	11
Directional Orientation, CP515B	392	....	....	....	....	....	06	....	....	....	36	....	22	07	....
Directional Plotting, CE455A (Total Rights)	354	−03	....	....	....	....	....	....	....	....	....	....	44	....	....
Directional Plotting, CE455A (Total Wrongs)	354	41	....	....	....	....	....	....	....	....	....	....	08	....	....
Driving Skill, CI307AX1	202	....	....	....	....	21	....	....	46	....	....	....	17	08	....
Figure Analogies, CI212AX1	468	....	08	04	34	....	11	....	−05	....	....	....	20	17	....
Figure Classification, CI213AX1	202	....	....	....	38	....	....	....	....	....	....	....	05	01	....
Flags, Figures, Cards, CP512A	392	....	....	....	....	....	06	....	....	....	15	....	−05	31	....
Flight Formations, CI654AX5	266	....	46	14	21	....	13	....	−02	....	....	....	00	11	....
Following Directions, CP402A	206	....	11	54	10	....	05	....	−03	....	....	....	25	09	....
Following Oral Directions, CI651AX3	266	....	18	25	18	....	09	....	07	....	....	....	21	08	....
Forced Landings, CI652AX4	266	....	16	38	11	....	−05	....	03	....	....	....	−02	07	....
General Information, CE505D (Navigator)	3000	....	....	....	....	....	....	27	02	....	....	....	20	10	20
General Information, CE505D (Pilot)	3000	....	....	....	....	....	....	00	30	....	....	....	−10	23	38
General Information (Tech.-Vocab.-Bomb.), CE505C	3000	....	....	....	....	....	....	....	15	....	....	....	04	08	33
General Information, (Tech.-Vocab.-Nav.), CE505C	3638	...	03	03	01	06	14	....	10	....	....	....	14	10	00
General Information (Tech.-Vocab.-Pil.), CE505C	3000	....	....	....	....	....	....	....	39	....	....	....	−08	17	34

and validities for printed tests grouped alphabetically[1]

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	A^2	r_{11}	Validity Nv	Pil.	Bom.	Nav.
....	....	....	....	05	....	....	24	−06	....	....	....	18	35	....	598	[3]18	....	....
....	04	....	....	00	....	....	−09	....	....	05	01	....	25	49	9600	02	....	....
															2847	....	....	23
....	22	....	....	−07	....	....	−05	....	....	−15	−07	....	46	88	7924	30	....	....
															1837	....	....	19
....	....	....	....	20	....	....	28	21	....	....	....	11	51	....	640	17	....	....
....	10	....	17	20	....	....	16	....	....	....	23	08	50	89	....	....	....	....
....	06	....	20	33	....	....	17	....	....	....	31	−01	76	87	....	....	....	....
17	....	....	....	36	15	01	−01	....	....	....	05	19	48	68	682	[3]19	....	....
....	−02	17	....	19	....	....	33	....	32	....	....	−02	55	56	....	....	....	....
....	−03	−07	....	03	....	....	13	....	03	....	....	09	37	46	....	....	....	....
....	10	....	−06	26	....	....	41	25	....	....	14	20	53	68	2155	[3]23	....	....
....	....	....	....	36	30	37	19	....	....	....	01	01	54	72	1520	13	....	....
....	....	....	....	....	....	....	32	....	....	10	14	....	26	....	1529	13	....	....
															200	....	....	24
....	....	....	....	24	....	....	41	....	....	....	....	−13	59	76	8630	16	....	....
															1048	....	21	....
....	02	14	....	16	....	....	42	....	....	−02	10	07	65	87	10925	22	....	....
															1820	....	18	....
															2701	....	....	55
....	....	....	....	31	....	....	41	08	....	....	....	26	56	74	931	26	....	....
															751	....	....	38
....	−06	26	....	17	....	....	30	....	42	....	....	45	76	79	....	....	....	....
....	−01	−03	....	02	....	....	09	....	−04	....	....	56	50	56	....	....	....	....
15	....	....	....	11	....	....	15	....	....	....	−05	42	53	55	820	[3]32	....	....
....	17	....	07	34	40	31	14	....	....	....	23	28	76	82	1009	19	....	....
															1675	....	....	61
....	....	....	....	03	16	32	06	....	....	....	15	04	30	78	2797	00	....	....
....	....	....	....	16	....	....	43	42	....	....	....	14	54	55	1950	[3]24	....	....
....	01	....	02	22	....	....	22	....	....	....	19	04	45	84	1302	23	....	....
....	10	....	07	16	....	....	02	....	....	....	26	17	54	75	5163	16	....	....
....	−06	....	−07	27	....	....	28	....	....	....	18	20	42	70	4797	[3]24	....	....
															1682	....	....	35
....	07	....	02	53	....	....	14	....	....	....	18	−05	53	....	1310	[3]08	....	....
....	−07	....	....	13	....	....	−02	....	....	....	60	....	55	....	12043	09	....	....
....	08	....	....	−05	....	....	−04	....	....	....	35	....	43	87	12043	24	....	....
....	00	12	....	01	....	....	−05	....	....	....	44	15	37	87	3151	12	....	....
															1829	....	04	....
															2701	....	....	15
−05	05	01	03	07	18	16	07	....	....	....	75	14	72	77	3151	09	....	....
															1829	....	05	....
															2701	....	....	33
....	−01	05	....	00	....	....	−01	....	....	....	41	01	47	80	3151	21	....	....
															1829	....	00	....
															2701	....	....	11

Test and Code No.	N_t	Ca	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Geography	1900	....	....	....	....	....	....	01	12	....	....	....	....	15	....
Gottschaldt Figures, QP901A-III	594	....	....	....	08	....	02	....	....	....	15	....	22	16	....
Hands, CP512A	658	....	1	−02	06	....	01	....	13	....	−05	....	17	13	....
History, AS153	1900	....	....	....	....	....	....	08	−08	....	....	....	....	08	....
Instrument Comprehension I, CI615A	468	....	−07	21	08	....	02	....	07	....	....	....	10	18	....
Instrument Comprehension II, CI616B	468	....		11	−02	....	13	03	03	....	....	....	09	17	....
Judgment of Proportions, CP206B	562	....	....	....	−09	06	15	....	02	....	13	....	02	22	....
Log Book Accuracy	236	....	06	06	09	....	02	....	−06	....	....	....	32	13	....
Map Distance, CP026B	658	....	01	....	−09	....	30	....	17	....	04	....	09	01	....
Map Memory, CI505AX1	179	....	....	....	....	....	....	....	....	09	54	....	....	35	....
Map Memory, CI505BX1	238	....	....	....	....	....	....	....	....	41	52	05	....	22	....
Map Memory, CI505AX2	238	....	....	....	....	....	....	....	....	14	58	07	....	35	....
Map Memory, CI505AX3	176	....	....	....	....	....	....	....	....	15	55	....	....	18	....
Map Planning, CI412AX1	170	....	....	....	−01	16	....	....	−08	....	....	....	12	45	....
Marking Accuracy	206	....	−07	01	−07	....	−10	....	−04	....	....	....	02	35	....
Mathematics A, CI702E	3000	....	....	....	....	....	....	....	−04	....	....	....	42	07	−12
Mathematics A, CI702F	3000	....	....	....	....	....	....	37	07	....	....	....	51	06	00
Mathematics B, CI206B	570	....	....	....	....	....	15	....	−04	09	16	14	....	00	....
Mathematics B, CI206C	3266	13	04	00	16	....	−01	12	15	....	....	....	48	−01	−07
Mathematics B, CI206B, CI700A	3966	....	....	....	07	04	04	....	08	....	13	....	57	−01	−02
Mechanical Functions, CI907A	153	....	....	....	....	05	05	....	42	....	....	....	....	35	....
Mechanical Information, CI905A	3791	....	01	−09	−03	00	07	....	74	....	....	....	−08	00	09
Mechanical Movements, CI004A	153	....	....	....	....	....	−04	....	38	....	....	....	....	37	....
Mechanical Principles, CI903A	7385	....	13	−05	14	03	−04	04	60	....	02	....	−07	08	−02
Mechanical Principles, CI903B	354	17	....	....	....	....	....	−12	58	....	02	....	−13	01	....
Memory for Landmarks, CI510AX1	417	....	....	....	....	....	....	....	....	61	20	44	....	16	....
Memory for Plane-Names, CI506AX1	238	....	....	....	....	....	....	....	....	58	02	51	....	29	....
Memory for Plane Silhouettes, CI503AX1	417	....	....	....	....	....	....	....	....	56	06	−09	....	34	....
Memory for Ships, CI504AX1	238	....	....	....	....	....	....	....	....	50	06	20	....	29	....
Memory for Tactical Plans, CI509AX	179	....	....	....	....	....	....	....	....	10	−12	....	....	−02	....
Nearest Point-Point Distance, CP607B	545	....	....	....	....	....	43	....	11	....	23	....	−04	09	....
Number Series Completion, CI215AX1	202	....	....	....	20	....	....	....	....	....	....	....	47	05	....
Numerical Operations, CI701B (Back)	6266	....	07	08	04	....	−03	−05	−03	....	....	....	81	04	05
Numerical Operations, CI701B (Front)	0206	04	01	02	−09	....	11	−06	−09	....	....	....	78	10	02
Numerical Operations, CI701B (Total)	372	....	....	....	02	14	....	....	−08	....	....	....	66	08	....

(Continued)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	h^2	r_{11}	Validity Nv	Pil.	Bom.	Nav.
....	....	....	....	....	....	....	06	....	....	58	61	....	75	....	4988	03	....	....
....	....	....	....	16	39	12	04	−02	....	....	−05	36	43	....	792	24	....	....
....	−07	....	06	02	....	....	17	46	....	....	07	10	35	95	1967	[a]20	....	....
....	....	....	....	....	....	....	−02	....	....	52	63	....	60	....	5014	−09	....	....
....	−08	....	24	21	34	00	44	....	....	....	22	07	57	68	9284	[a]20	....	....
....	−07	....	14	09	36	04	53	....	....	02	24	25	65	84	8889	[a]32	....	....
30	....	....	....	08	....	....	11	03	....	....	22	29	35	52	765	[a]08	....	....
....	02	....	60	03	....	....	19	....	....	....	09	−11	56	75	....	....	....	....
....	14	....	−06	06	....	....	06	08	....	....	19	38	35	72	2752	[a]20	....	....
															911	....	....	34
....	....	....	....	06	....	....	16	....	....	....	42	26	70	84	793	17	....	....
....	....	....	....	23	....	....	08	....	....	....	05	....	56	66	1083	19	....	....
....	....	....	....	17	....	....	16	....	....	....	23	....	59	83	1312	17	....	....
															1577	....	....	34
....	....	....	....	14	....	....	21	....	....	....	31	31	61	67	176	16	....	....
24	....	....	....	31	....	....	27	....	....	....	−08	28	57	78	1530	21	....	....
....	−06	....	50	06	....	....	−04	....	....	....	05	−03	41	86	....	....	....	....
....	06	01	....	12	....	....	16	....	....	....	53	33	63	92	3151	12	....	....
															1829	....	14	....
															2701	....	....	42
....	−07	....	....	24	....	....	07	....	....	16	37	....	64	....	8840	13	....	....
....	....	....	....	57	....	....	13	....	....	....	20	10	51	....	12057	09	....	....
															1378	....	13	....
....	−03	−06	04	47	....	....	18	....	02	14	27	19	72	84	18657	10	....	....
10	−03	02	....	40	−04	00	13	−04	....	....	20	22	68	80	3151	10	....	....
															1829	....	13	....
															2701	....	....	40
....	....	....	....	22	....	24	−03	....	....	....	24	04	41	80	932	[a]29	....	....
10	03	−01	−11	04	....	....	04	....	....	....	11	15	65	85	3151	28	....	....
															1829	....	00	....
															2701	....	....	06
....	....	....	....	25	....	....	28	....	....	....	05	51	69	76	2572	[a]23	....	....
13	09	−02	−10	20	01	00	22	04	....	....	20	51	84	76	10025	37	....	....
															1829	....	10	....
															2701	....	....	25
....	18	−02	....	34	....	....	12	....	28	03	03	54	93	....	9930	[a]33	....	....
....	....	....	....	15	....	....	07	....	....	....	07	−13	68	82	1421	[a]10	....	....
....	....	....	....	10	....	....	06	....	....	....	13	....	71	82	1771	22	....	....
															1652	....	....	21
....	....	....	....	−06	....	....	38	....	....	....	16	02	61	82	873	21	....	....
....	....	....	....	10	....	....	31	....	....	....	−08	....	51	66	1694	17	....	....
....	....	....	....	09	....	....	08	....	....	....	57	32	47	68	1564	[a]19	....	....
....	....	....	....	21	....	....	18	05	....	....	−12	16	38	62	4045	19	....	....
															608	....	00	....
															600	....	....	15
....	....	....	....	36	19	00	04	....	....	....	10	16	47	78	2309	[a]13	....	....
....	03	14	10	11	....	....	05	....	....	....	06	05	73	79	3151	03	....	....
															1829	....	12	....
															2701	....	....	41
....	08	22	23	02	....	....	−01	....	−06	....	05	−10	78	64	3151	−01	....	....
															1829	....	12	....
															2701	....	....	38
03	....	....	....	16	11	07	07	....	....	....	09	−08	[illegible]	81	....	....	....	....

Test and Code No.	N	Ca	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Organizational Planning, CI407AX1	202	...	...	...	...	13	...	...	20	...	...	...	41	18	...
Organizational Planning, CI407BX1	266	...	02	35	28	...	05	...	20	...	...	...	38	16	...
Path Length, CP628B	392	...	...	...	...	...	25	...	...	...	−11	...	08	10	...
Pattern Analysis, CP512A	392	...	...	...	...	...	04	...	...	...	35	...	26	26	...
Pattern Assembly, CP504A	202	...	...	...	...	...	52	...	02	...	...	...	...	31	...
Pattern Comprehension, CP503A	202	...	...	...	13	...	...	...	...	...	...	...	−04	24	...
Pattern Comprehension, CP503AX1	153	...	...	...	...	...	00	...	06	...	...	...	...	29	...
Physics, CI801A	153	...	...	...	...	...	01	...	21	...	...	...	...	17	...
Physical Principles, CI801BX	1900	...	...	...	...	...	...	−11	51	...	...	...	...	−03	...
Picture Integration, CP104A	392	...	...	...	...	...	−08	...	...	...	28	...	16	27	...
Planning Air Maneuvers, CI408AX1	202	...	...	...	...	07	...	...	17	...	...	...	16	00	...
Planning Air Maneuvers, CI408AX3	635	...	01	06	43	−11	−19	...	20	...	...	...	08	06	...
Planning A Circuit, CI401A	202	...	...	...	...	19	...	...	24	...	...	...	02	41	...
Planning A Course, CI406AX2	436	...	06	17	41	10	16	...	00	...	...	...	20	05	...
Plotting Accuracy, CE453A (Total Rights)	354	19	...	...	...	...	...	...	...	...	...	...	60	...	...
Plotting Accuracy, CE453A (Total Wrongs)	354	51	...	...	...	...	...	...	...	...	...	...	22	...	...
Plotting Test, CE452A. (Total Rights)	354	22	...	...	...	...	...	...	...	...	...	...	51	...	...
Plotting Test, CE452A. (Total Wrongs)	354	50	...	...	...	...	...	...	...	...	...	...	00	...	...
Practical Estimations I, CI308AX1	170	...	...	...	01	36	...	...	32	...	...	...	13	−05	...
Practical Estimations II, CI308AX1	170	...	...	...	11	02	...	...	32	...	...	...	28	00	...
Practical Judgment (Mechanical Items), CI301BX1	202	...	...	...	...	36	...	...	54	...	...	...	07	20	...
Practical Judgment (Non-Mechanical), CI301BX1	202	...	...	...	...	39	...	...	13	...	...	...	18	16	...
Practical Judgment I (Non-Mechanical), CI301BX3	170	...	...	...	−07	30	...	...	21	...	...	...	20	02	...
Practical Judgment II (Work Plan), CI301BX3	170	...	...	...	01	45	...	...	32	...	...	...	18	−08	...
Pursuit-Path Tracing, CP512A	796	...	...	...	17	−01	...	...	07	...	11	...	25	81	...
Reading Comprehension, CI614G	6372	...	...	...	07	04	...	05	37	...	...	...	12	03	04
Reading Comprehension, CI614H	266	10	16	26	22	...	−01	05	04	...	...	...	12	−03	...
Route Planning, CI411AX1	436	...	17	15	37	11	−02	...	06	...	...	...	02	17	...
Sequence of Maneuvers, CI410A	202	...	...	...	...	38	...	...	00	...	...	...	30	00	...
Shorter Line–Line Lengths, CP606B	545	...	...	...	...	...	44	...	04	...	11	...	09	10	...
Shorter Path–Path Distance, CP608B	545	...	...	...	...	...	46	...	−01	...	−11	...	17	28	...
Signal Interpretation, CI036AX2	266	...	50	06	30	...	10	...	−02	...	...	...	05	06	...
Spatial Orientation I, CP501A	153	...	...	...	...	...	02	...	02	...	...	...	...	62	...
Spatial Orientation I, CP501B	7802	...	−02	12	17	03	13	05	01	12	20	00	18	83	18
Spatial Orientation II, CP503B	6800	...	...	...	...	...	...	08	18	09	06	06	−03	54	09

(Continued)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_2	h^2	r_{11}	Validity: Nv	Pil.	Bom.	Nav.
05	....	....	....	39	....	....	29	....	....	....	27	02	50	48	102	25	....	....
....	00	....	06	00	....	....	15	....	....	....	14	−08	46	....	....	....	....	....
....	....	....	....	21	....	....	23	17	....	....	....	19	27	25	2491	[a]23	....	....
....	....	....	....	15	....	....	15	12	....	....	....	18	35	37	640	16	....	....
....	....	....	....	14	....	....	01	....	....	....	03	16	42	59	320	[a]18	....	....
....	....	....	....	33	24	13	06	....	....	....	18	50	55	....	1081	10	....	....
....	....	....	....	45	....	....	−01	....	....	....	21	28	41	....	523	09	....	....
....	....	....	....	17	....	....	03	....	....	....	68	25	63	65	6079	[a]10	....	....
....	....	....	....	....	....	....	17	....	....	11	38	....	46	....	6079	10	....	....
....	....	....	....	42	....	....	40	−09	....	....	....	26	61	77	309	28	....	....
51	....	....	....	30	....	....	23	....	....	....	26	10	54	61	374	[a]27	....	....
46	13	....	−06	18	01	−01	32	....	....	....	17	11	69	73	2279	18	....	....
40	....	....	....	10	....	....	29	....	....	....	24	36	57	83	3578	[a]26	....	....
−06	00	....	11	24	....	....	45	....	....	....	16	10	64	81	877	[a]17	....	....
....	15	18	....	02	....	....	30	....	38	....	....	03	75	78	....	....	....	....
....	−08	04	....	07	....	....	08	....	01	....	....	−02	33	72	....	....	....	....
....	15	20	....	00	....	....	25	....	46	....	....	09	65	80	....	....	....	....
....	−09	01	....	10	....	....	13	....	03	....	....	−05	39	77	....	....	....	....
36	....	....	....	−04	....	....	00	....	....	....	01	07	39	42	1283	[a]14	....	....
31	....	....	....	−08	....	....	02	....	....	....	03	22	35	....	1283	[a]12	....	....
0	....	....	....	18	....	....	05	....	....	....	12	29	50	62	2487	[a]18	....	....
36	....	....	....	12	....	....	01	....	....	....	17	30	48	62	4780	[a]13	....	....
28	....	....	....	03	....	....	34	....	....	....	30	29	49	54	....	....	....	....
06	....	....	....	40	....	....	10	....	....	....	03	04	51	41	....	....	....	....
27	....	....	....	−11	−08	06	17	....	....	....	10	−03	50	66	1148	06	....	....
18	−02	10	....	27	27	03	15	....	....	....	60	24	79	76	10925 1829 2701	20	.... 13	 38
....	00	16	07	19	....	....	−02	....	06	10	60	30	68	85	4821	17	....	....
47	−10	....	−03	22	....	....	17	....	....	....	−04	29	63	77	1953	[a]16	....	....
35	....	....	....	21	....	....	19	....	....	....	30	00	80	66	247	00	....	....
....	....	....	....	13	....	....	06	−04	....	....	15	03	27	....	3616 546 333	[a]17	.... [a]10	.... [a]13
....	....	....	....	13	....	....	22	....	....	....	10	28	52	60	2721 603 643	24	.... [a]03	.... [a]28
....	11	....	18	41	....	....	06	....	....	....	17	01	69	77	2112	[a]21	....	....
....	....	....	....	12	....	....	13	....	....	....	16	24	30	91	2963 412	17	....	 36
06	07	16	21	12	09	04	10	....	....	60	04	09	60	63	10925 1829 2701	20	.... 13	 38
....	−01	03	....	17	....	....	16	....	....	20	14	25	33	60	10925 1829 2701	28	.... 06	 39

Test and Code No.	N_t	C_s	I_1	I_2	I_3	J	LE	MB	ME	M_1 PM	M_2 VM	M_3	N	P	PI
Spatial Reasoning, CI211BX1	404	....	....	....	38	14	....	....	02	....	....	....	18	16	....
Spatial Visualization I, CI204AX1	202	....	....	....	18	....	....	....	....	....	....	....	03	14	....
Spatial Visualization I, CI204AX2	266	....	01	03	34	....	06	....	04	....	....	....	05	20	....
Spatial Visualization II, CI203AX1	202	....	....	....	17	....	....	....	....	....	....	....	17	08	....
Speed of Identification (Non-Rotated), CP610A	392	....	....	....	....	....	....	....	....	....	−14	....	−06	70	....
Speed of Identification (Rotated), CP610A	7649	00	04	02	12	10	03	01	09	05	09	09	04	64	08
Table Reading, CP621A	392	....	....	....	....	....	09	....	....	....	10	....	50	30	....
Tool Function, CI906A	153	....	....	....	....	....	11	....	77	....	....	....	....	30	....
Vocabulary, CI604B	1903	....	....	....	....	....	....	04	−11	....	....	....	....	11	....

[1] Decimal points have been omitted.
[2] Validities, and other statistics, are weighted averages. See chapter 28 for explanation.
[3] Derived from combinations of data on similar forms.

(Continued)

P1	PM_1 PC	PM_2 PP	PM_3 PS	R_1 GR	R_2	R_3	S_1 SR	S_2	S_3	SS	V	V_f	h^2	r_{11}	Validity Nv	Pil.	Bom.	Nav.
12	....	....	....	45	05	38	26	....	....	....	20	16	72	86	999 1291	11	....	 48
....	....	....	....	34	32	21	24	....	....	....	15	56	71	85	196	26	....	....
....	−04	....	05	39	....	....	14	....	....	....	26	53	69	....	1550 735	12	....	 45
....	....	....	....	44	25	36	17	....	....	....	18	42	75	87	3088 594	17	....	 36
....	....	....	....	−06	....	....	10	....	....	....	....	03	52	....	....	....	....	....
03	07	11	05	−01	31	03	05	−01	13	....	05	10	63	76	10925 1829 2701	18	.... 06	 18
....	....	....	....	13	....	....	35	−11	....	....	....	−14	53	84	9602 1021	18	 24	....
....	....	....	....	05	....	....	18 04	....	....	.... −01	−02 71	−11	74 53	....	78 1662	17 −08	....	....

APPENDIX C

Intercorrelations, Means, and Standard Deviations of 65 Selected Tests in Samples of Unclassified Aviation Students at Sheppard Field, Texas[1]

[1] The number of examinees varied from test to test. The following table shows the number of examinees on which each correlation coefficient that was computed was based. No self correlations (that is, reliability coefficients) were computed. Decimal points that properly precede entries in the matrix have been omitted.

Product-moment correlation coefficients among

Test		Test	N
1- 8	and	1- 8	1,588
1- 8	and	9-16	440
1- 8	and	17-25	385
1- 8	and	26	180
1- 8	and	27-35	417
1- 8	and	36-44	431
1- 8	and	45-65	1,588
9-16	and	9-16	1,568
9-16	and	17-26	412
9-16	and	27-35	407
9-16	and	36-44	392
9-16	and	45-65	1,567
17-26	and	17-26	1,539
17-26	and	27-35	415
17-26	and	36-44	402
17-26	and	45-65	1,539
27-35	and	27-35	1,502
27-35	and	36-44	423
27-35	and	45-65	1,502
36-44	and	36-44	1,567
36-44	and	45-65	1,567
45-65	and	45-65	8,158

The following table shows the number of testees on which the computation of means and standard deviations was based.

Tests	N
1- 8	1,588
9-16	1,567
17-25	1,539
26	1,364
27-35	1,502
36-44	1,567
45-65	8,158

[3]For Tests 45-65, the scoring formulas used are presented in Table 3.18 of report no. 2 in the series of AAF aviation psychology research reports. *Vid.* DuBois, P.H. ed. *The Classification Program*, AAF Aviation Psychology Research Reports, No. 2, Washington: Government Printing Office, 1947.

For other tests for which the scoring formula is not specified in footnote 3, the scoring formula included an appropriate correction for chance success.

[4]Score was number of items marked correctly.

A

Test No.			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	Map Memory	CI505BX	....	39	[illegible]	25	30	25	20	29	28	31	29	29	23	33	29	-05	36	28	39	36	32	13	22	39	31	34	40	39	[illegible]
2	Figure Analogies	CI212AX1	39	....	59	34	35	25	22	33	27	42	47	40	30	46	46	07	31	47	53	35	47	22	21	42	35	33	28	46	37
3	Spatial Visual II	CI203A	35	60	....	36	30	23	22	26	34	42	61	40	38	41	54	07	25	41	63	34	50	24	19	42	37	42	21	44	41
4	Plan Air Maneuvers	CI406AX2	25	34	34	....	15	13	13	16	24	28	35	29	26	33	38	03	25	17	33	28	29	18	15	27	25	19	13	25	25
5	Vocabulary	CI604BX1[1]	30	35	30	15	....	19	17	26	22	25	31	14	10	15	25	-08	17	31	23	30	32	06	15	23	26	18	15	29	11
6	Map Distance	CP626B[1]	25	25	23	13	19	....	40	46	14	29	19	22	26	27	15	20	26	24	29	29	27	26	27	25	27	26	15	30	23
7	Estimation of Length	CP631B[1]	20	22	22	13	17	40	....	30	09	23	10	15	17	21	09	23	10	12	14	23	14	23	24	13	19	23	16	27	21
8	Speed of Identification	CP610C	29	33	26	16	26	46	30	....	24	37	26	33	30	39	23	19	28	29	23	38	21	21	31	17	26	24	17	43	25
9	Memory for Plane Silhouettes	CI503AX1	28	27	34	24	22	14	09	24	....	31	43	33	24	25	38	11	31	15	45	21	36	23	21	45	29	29	50	30	30
10	Directional Orientation	CP515F	31	42	42	28	25	29	23	37	31	....	48	41	37	38	43	17	33	23	47	28	42	31	29	41	41	40	28	46	39
11	Visualization of Maneuvers	CI467CX2	29	47	61	35	31	18	10	26	43	48	....	38	34	30	57	14	28	27	48	33	48	27	31	60	42	38	34	37	34
12	Planning a Circuit	QP901A-I	29	40	40	29	14	22	15	33	38	41	38	....	39	48	44	10	28	28	47	31	31	37	23	42	33	58	25	36	43
13	Path Tracing	QP901A-V	23	30	38	26	10	26	17	30	24	37	34	39	....	33	32	20	16	16	38	31	25	31	23	29	25	34	21	31	29
14	Maze Tracing	QP901A-VII	33	46	41	33	15	27	21	39	25	38	30	48	33	....	41	09	29	23	43	44	34	22	29	32	31	33	24	41	42
15	Formation Visualization	CP814A	29	48	54	38	25	15	09	23	28	43	57	44	32	41	....	14	29	30	57	30	49	34	32	57	30	38	30	39	34
16	Objectivity of Perception	CP806CX2[1]	-05	07	07	03	-08	20	23	19	11	17	14	10	20	00	14	....	15	08	10	14	21	40	30	10	17	34	12	12	10
17	Visual Memory	CI514A	36	31	25	25	17	26	10	28	31	33	28	28	16	29	29	15	....	17	31	33	22	14	16	34	23	22	32	23	17
18	Figure Classification	CI213[illegible]X1	28	47	41	17	31	24	12	29	15	23	27	25	16	23	30	08	17	....	33	17	32	18	14	27	23	25	16	23	17
19	Spatial Visualization I	CI204AX1	39	56	63	33	23	29	14	23	45	47	48	47	38	43	57	19	31	33	....	33	47	30	26	43	38	41	19	40	35
20	Map Planning	CI412AX1	36	35	34	28	20	29	23	33	21	23	33	31	21	44	30	14	23	17	32	....	28	34	28	25	32	33	21	36	38
21	Object Recognition	CP823A	32	47	50	29	32	27	14	21	26	43	48	31	25	34	49	21	22	22	47	28	....	22	30	38	39	40	-16	41	29
22	Objectivity of Perception	CP806CX1[1]	13	22	24	18	06	26	23	21	23	31	27	37	31	22	34	40	14	18	30	24	22	....	36	22	22	27	29	33	27
23	Position Orientation	CP528B	22	21	19	18	15	27	24	31	21	29	31	23	23	29	32	29	18	14	36	28	30	36	....	30	38	41	25	29	21
24	Aerial Orientation	CP520C	39	42	42	27	23	25	13	17	45	41	60	42	29	32	57	10	24	27	43	25	38	22	30	....	32	32	23	26	28
25	Object Identification I	CP521A-I	31	35	37	25	26	27	19	25	29	41	42	23	25	31	39	17	23	23	38	32	30	22	38	32	....	62	25	33	25
26	Object Identification II	CP521A-II	34	33	42	19	18	26	23	24	[illegible]	40	38	38	34	33	34	24	22	25	41	33	40	37	41	32	62	....	15	32	27
27	Plane Position Memory	CI512A	40	26	21	13	15	15	16	17	[illegible]	20	34	25	21	24	30	12	32	18	19	21	16	29	25	23	25	16	....	28	25
28	Decoding	CI214AX2	36	46	44	25	29	30	27	43	[illegible]	48	37	36	31	41	39	12	23	28	40	26	41	23	39	26	33	32	28	....	30
29	Route Planning	CI411AX1	26	37	41	28	11	23	21	25	30	30	34	43	29	42	34	10	17	17	35	36	29	27	21	28	25	27	25	36	....
30	Flight Formation	CI554AX5	30	45	22	24	28	20	23	35	23	22	23	37	25	29	30	08	21	36	31	26	32	25	26	27	22	26	25	41	36
31	Aerial Landmarks	CP828C	36	41	42	29	14	11	13	19	39	46	62	28	38	29	49	13	35	21	46	24	34	38	32	41	35	32	40	34	36
32	Pattern Assembly	CP804A	25	36	32	19	13	36	33	33	28	30	34	18	24	22	23	31	21	23	31	24	19	28	22	17	28	41	23	22	21
33	Block Counting	CP512B-4	36	40	52	34	16	42	38	47	32	41	42	45	37	45	47	26	31	23	46	45	34	28	41	30	37	38	27	30	43
34	Discrimination Reaction Time I and II	CP634A	22	19	18	08	17	24	14	32	18	32	23	19	27	22	20	11	16	14	15	20	19	26	26	16	20	25	24	31	27
35	Discrimination Reaction Time III & IV	CP634A	23	24	20	10	23	23	27	39	17	30	37	26	25	23	25	12	21	11	19	31	30	25	32	23	27	23	21	33	27
36	Plane Name Memory	CI503AX2	36	31	27	07	24	30	19	18	34	14	14	15	18	09	13	17	22	23	32	18	38	20	17	18	20	11	40	20	10
37	Planning a Course	CI406AX3	34	46	29	27	30	17	27	35	17	30	35	23	18	37	23	13	22	30	28	35	34	33	20	29	37	32	17	32	30
38	Compass Orientation	CI600A[1]	35	36	30	09	31	24	23	43	08	14	11	-07	08	00	15	00	15	25	28	38	38	27	44	31	33	35	30	32	27
39	Competitive Planning	CI409AX2	23	34	33	23	15	14	16	20	18	21	20	21	14	31	24	18	19	25	31	34	33	16	18	27	29	28	13	27	27
40	Camouflaged Outlines	CP821A	22	41	42	21	34	29	30	35	28	30	43	30	33	36	36	29	26	26	40	28	32	37	33	26	30	35	19	31	32
41	Angle Estimation	CP318A	20	22	35	18	19	-01	-02	12	21	28	42	22	19	21	34	12	19	25	44	06	31	19	15	39	20	18	23	16	31
42	Spatial Reasoning	CI211BX2	32	48	39	28	34	36	25	30	19	40	36	36	25	29	38	17	28	33	37	43	37	34	30	35	39	40	15	37	30
43	Tool Function	CI906B	11	11	24	06	04	05	00	07	21	13	28	25	14	14	25	05	04	16	38	-08	17	07	09	12	15	13	10	04	15
44	AAF Word Knowledge	CI604C	22	34	37	18	79	10	12	18	07	12	24	-02	02	06	16	05	18	41	26	13	38	17	07	32	24	21	19	31	24
45	Dial & Table Reading	CP621-622A	39	49	40	24	35	36	33	51	26	46	37	36	30	38	36	22	27	31	38	40	42	34	41	35	36	40	28	49	33
46	Biographical Data (P)	CE602D	05	06	15	09	-04	06	05	03	14	06	16	19	09	09	16	03	11	07	25	06	13	11	03	31	08	07	09	01	15
47	Biographical Data (N)	CE603D	06	10	06	01	07	07	03	09	05	06	06	06	05	06	04	06	07	08	06	10	10	12	13	09	11	09	08	14	11
48	Spatial Orientation I	CP501B	28	30	30	18	21	37	32	47	29	34	31	32	28	33	28	19	30	20	32	30	28	25	32	38	28	31	25	37	20
49	Spatial Orientation II	CP503B	34	42	48	27	27	21	18	24	30	41	47	60	30	38	45	10	[illegible]	27	52	23	38	20	20	42	25	28	34	23	33
50	Reading Comprehension	CI614	31	49	45	23	55	17	14	25	25	32	40	25	16	21	35	04	21	38	34	16	41	18	15	34	23	24	28	36	25
51	Instrument Comprehension	CI616C	28	31	34	22	24	23	20	21	38	36	46	31	22	25	43	19	22	27	43	23	33	24	30	51	31	31	27	27	25
52	Mechanical Principles	CI903B	18	38	49	27	12	08	05	02	28	37	50	26	29	20	49	12	15	25	43	17	39	22	17	42	29	30	19	16	31
53	Speed of Identification	CP610A	26	26	27	14	21	32	28	42	37	31	30	36	21	23	29	12	28	18	33	36	30	18	22	34	35	27	34	30	21
54	Numerical Operations I	CI701B	16	25	19	07	33	26	26	30	04	24	14	08	12	13	10	15	13	17	12	28	22	23	31	11	25	29	14	33	12
55	Numerical Operations II	CI701B	22	34	26	13	34	26	26	38	06	36	18	14	13	16	17	18	13	21	30	27	29	22	35	16	25	29	15	37	17
56	Mechanical Information	CI905B	06	15	25	14	03	02	03	-03	16	15	25	24	11	10	30	-01	05	14	29	-01	12	07	02	30	09	09	08	01	17
57	General Information	CE505F	10	13	21	13	27	08	13	03	38	19	27	22	10	11	24	-02	10	14	26	08	11	15	10	33	14	08	22	05	15
58	Practical Judgment	CI301C	23	34	35	20	33	06	09	12	15	23	25	22	11	20	26	04	13	25	27	14	28	12	11	28	16	14	16	22	21
59	Arithmetic Reasoning	CI706C	28	50	53	26	35	19	18	24	15	36	34	26	22	26	35	15	18	35	38	23	40	21	22	25	26	28	16	40	27
60	Rotary Pursuit	CP410B	07	06	11	10	06	15	11	14	18	14	19	18	21	12	14	12	16	11	20	15	13	13	17	23	15	18	16	03	10
61	Rudder Control	CM120D	04	09	16	10	-04	03	02	01	11	16	19	16	16	14	17	07	05	04	20	13	12	09	12	26	06	07	10	-01	14
62	Finger Dexterity	CM116A	13	16	17	19	08	19	22	27	11	12	14	16	11	14	14	14	17	15	23	22	30	16	23	17	30	27	13	14	08
63	Complex Coordination	CM701A	21	31	30	20	15	22	20	32	29	32	36	36	33	28	34	12	20	23	35	23	22	24	32	40	33	33	26	21	30
64	Two-Hand Coordination	CM101A	09	17	19	10	-01	14	15	13	19	17	22	25	20	17	21	10	08	15	24	16	20	19	30	23	.19	24	12	19	24
65	Discrimination Reaction Time	CP[illegible]	27	27	36	20	27	[illegible]	19	33	29	35	41	32	24	31	[illegible]	14	17	31	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

TEST NUMBER

B

18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
26	30	36	32	13	22	39	31	34	40	36	26	30	35	25	38	22	23	36	34	35	23	22	20	32	11	22	39	05	06	29	34	31	25	18	26	19	22	06	16
47	55	35	47	22	21	42	35	33	28	46	37	45	41	36	49	19	24	31	48	36	34	41	22	38	11	34	49	08	10	30	42	49	31	35	28	23	34	15	13
41	63	34	60	24	19	42	37	42	21	44	41	32	42	32	[illegible]	18	20	27	39	30	33	42	35	29	24	37	40	15	06	26	48	45	34	49	27	19	26	25	21
17	33	28	29	18	15	27	25	19	13	25	29	[illegible]	29	19	34	08	10	07	27	09	23	21	18	[illegible]	06	15	24	09	01	18	27	23	22	27	14	07	13	14	12
31	23	20	22	06	15	23	26	16	15	29	11	28	14	13	15	17	23	24	36	31	15	24	19	34	08	[illegible]	35	-04	07	21	27	55	24	12	21	33	34	05	27
24	29	29	27	26	27	25	27	26	15	30	23	20	11	36	42	24	23	20	17	24	14	29	-01	26	05	10	36	06	07	37	21	17	23	08	32	26	26	02	09
12	14	23	14	23	24	13	19	22	16	27	21	23	13	33	38	24	27	19	27	23	19	30	-02	25	00	12	33	05	03	32	18	14	20	05	28	26	26	03	12
29	33	38	21	21	31	17	25	24	17	43	25	25	19	33	47	32	38	26	35	43	20	35	12	39	07	18	51	03	09	47	24	25	21	02	42	39	38	-03	02
15	45	21	28	23	21	45	29	29	50	30	30	23	39	28	32	18	17	34	17	06	16	28	21	19	21	07	26	14	05	29	39	25	33	28	37	04	08	16	38
23	47	28	42	31	29	41	41	40	28	48	39	32	45	29	41	32	30	14	39	14	21	30	28	40	13	12	46	05	03	34	41	32	36	37	31	24	26	15	19
27	48	23	48	27	31	60	42	38	34	37	34	23	52	24	42	23	27	14	35	11	20	43	42	36	28	24	37	16	06	31	47	40	46	50	30	14	18	25	37
28	47	[illegible]	31	37	23	42	33	38	25	36	43	37	28	18	45	19	30	15	33	-07	21	30	22	36	25	-02	36	19	06	32	40	25	31	35	26	08	11	24	22
16	36	31	25	31	23	29	25	34	21	31	29	25	28	24	37	27	25	18	18	08	14	33	19	25	14	02	30	09	05	23	30	16	22	29	21	12	13	11	10
23	43	44	34	22	29	32	31	33	24	41	42	29	29	22	45	22	23	09	37	00	31	33	21	39	14	06	38	09	03	33	35	21	25	29	23	13	16	10	11
30	57	30	49	34	32	57	39	38	30	39	34	30	48	23	47	20	25	13	33	15	24	36	34	38	25	16	36	16	04	28	48	35	43	49	29	10	17	20	24
08	19	14	21	40	29	10	17	24	12	12	10	08	13	31	26	11	12	17	13	00	19	29	12	17	05	05	22	03	06	16	10	04	10	13	12	15	18	-01	-02
17	31	23	22	14	16	24	23	22	32	23	17	21	35	21	31	16	21	22	22	15	19	26	19	28	04	18	27	11	07	30	29	21	22	15	23	13	13	05	10
....	33	17	32	18	14	27	23	25	16	28	17	36	21	23	23	14	11	23	30	25	25	26	25	33	16	41	31	07	08	20	27	33	27	25	18	17	21	14	14
33	....	33	47	30	26	43	38	41	19	40	35	31	46	31	46	15	19	32	28	28	31	49	44	37	35	28	38	22	08	32	52	34	43	53	33	12	20	29	26
17	33	....	28	24	28	25	32	33	21	36	38	26	24	24	45	30	31	18	35	38	34	28	08	43	-03	13	40	05	10	30	25	16	23	17	26	23	28	-01	09
32	47	28	....	22	30	38	39	40	16	41	29	32	34	19	34	19	20	28	34	38	33	33	31	37	17	38	42	12	10	28	38	41	33	39	20	22	29	12	11
18	30	24	22	....	26	22	22	27	29	33	27	25	26	28	36	26	25	20	23	27	16	37	19	34	07	17	34	11	12	25	20	18	24	22	18	23	22	07	13
14	26	28	30	26	....	30	38	41	25	29	21	26	32	22	41	26	32	17	20	44	18	33	15	30	09	07	41	03	13	32	20	15	30	17	22	31	32	02	10
27	43	25	38	22	30	....	32	32	23	26	28	27	41	17	39	16	23	18	29	31	27	26	39	33	12	23	35	21	09	28	42	34	51	42	26	11	16	20	33
23	38	32	39	22	38	32	....	62	25	33	25	22	35	26	37	20	27	20	27	35	[illegible]	39	20	39	15	24	36	06	11	28	25	23	31	29	25	25	25	09	14
25	41	33	40	27	41	32	62	....	15	32	27	26	32	31	38	25	23	11	32	35	28	35	18	40	13	21	40	07	02	31	28	24	31	30	27	29	29	09	08
16	19	21	16	29	25	23	25	15	....	28	25	25	30	23	27	24	21	40	17	30	13	19	23	15	10	19	28	09	08	25	24	28	27	19	24	14	15	08	22
28	40	36	41	83	29	26	33	32	28	....	36	41	34	22	39	31	33	20	32	32	27	31	16	37	04	31	49	01	14	37	33	36	27	16	30	33	37	01	05
17	35	38	29	27	21	28	25	27	25	36	....	36	36	21	43	27	27	10	30	27	27	32	31	30	15	24	33	15	11	30	33	25	25	31	21	12	17	17	15
36	31	26	32	25	26	27	22	26	25	41	36	....	27	20	33	29	33	23	44	39	25	34	15	51	02	45	44	03	13	27	29	[illegible]	[illegible]	[illegible]	[illegible]	27	30	02	08
21	46	24	34	28	32	41	35	32	30	34	36	27	....	22	37	20	20	21	20	27	19	26	39	24	16	24	33	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
23	31	24	19	28	22	17	26	31	23	22	21	20	22	....	39	19	17	13	13	19	11	[illegible]	22	17	13	05	23	11	[illegible]	32	25	13	16	16	35	12	[illegible]	[illegible]	10
23	46	45	34	36	41	39	37	38	27	39	43	36	37	39	....	31	35	17	36	37	26	40	31	44	18	24	47	11	08	44	38	25	35	30	37	25	23	13	17
14	15	30	19	26	26	16	20	25	24	31	27	29	20	19	31	....	48	28	21	32	21	30	14	30	00	24	39	03	14	31	19	25	19	08	24	27	26	-04	02
11	19	31	20	25	32	23	27	22	21	33	27	33	20	17	35	48	....	15	35	34	26	25	12	32	02	26	41	04	14	30	19	25	21	09	22	28	27	00	08
23	32	18	28	20	17	18	20	11	40	20	10	23	21	13	17	28	15	....	17	26	13	25	15	19	-01	24	27	04	11	20	22	25	22	07	23	18	15	-03	13
30	28	35	34	23	20	29	27	32	17	32	30	44	20	13	36	21	35	17	....	30	29	29	12	42	-01	28	45	-01	07	23	28	35	26	18	18	31	37	04	06
25	28	38	38	27	44	31	33	35	30	32	27	39	27	19	37	32	34	26	30	....	11	17	07	33	-01	25	41	-03	08	22	24	27	28	11	25	34	35	01	08
25	31	34	33	16	18	27	29	28	13	27	27	25	19	11	26	21	26	13	29	11	....	23	16	29	04	18	31	-08	09	17	23	24	21	16	17	23	28	01	02
26	49	28	33	37	33	26	39	35	19	31	32	34	26	30	40	30	25	25	29	17	23	....	24	36	16	26	35	01	04	34	33	27	25	24	27	19	20	09	06
25	44	08	31	19	15	39	20	18	23	16	31	15	39	22	31	14	12	15	12	07	16	24	....	15	19	14	19	15	-01	20	35	24	31	36	21	02	06	18	24
33	37	43	37	34	30	33	39	40	15	37	30	51	24	17	44	30	32	19	42	33	29	36	15	....	04	28	52	-04	06	27	29	41	25	21	17	38	41	04	02
16	35	-03	17	07	09	12	15	13	10	04	15	02	16	13	18	00	02	-01	-01	-01	04	16	19	04	....	01	02	32	-09	04	19	10	17	41	06	-13	-07	60	31
41	28	13	38	17	07	23	24	21	19	31	24	45	24	05	24	24	26	24	28	25	18	26	14	28	01	....	32	-14	05	13	26	57	19	14	15	28	32	06	11
31	38	40	42	34	41	35	36	40	28	49	33	44	33	23	47	39	41	27	45	41	31	35	19	52	02	32	....	03	17	40	35	41	37	21	32	52	55	02	14
07	22	05	12	11	03	21	06	07	09	01	15	03	16	11	11	03	04	04	-01	-02	-03	01	15	-04	32	-14	03	....	20	08	13	01	19	27	13	-13	-11	35	37
08	08	10	10	12	13	09	11	09	08	14	11	13	01	06	08	14	14	11	07	03	09	04	-01	06	-09	08	17	20	....	09	04	13	08	00	08	14	13	-07	06
20	32	30	26	25	32	23	28	31	25	37	30	27	39	32	44	21	30	20	23	22	[illegible]	34	20	27	04	13	40	08	09	....	39	18	28	15	45	23	23	01	15
27	52	25	38	20	20	42	25	28	24	33	33	29	44	25	38	19	19	22	28	24	23	33	35	29	19	26	35	13	04	39	....	33	37	40	36	10	17	18	23
35	34	16	41	18	15	34	23	24	28	36	25	41	27	13	25	25	25	25	35	27	24	27	24	41	10	57	41	01	13	18	33	....	33	32	18	29	35	18	27
27	43	23	33	24	30	51	31	31	27	27	25	28	34	16	35	19	21	22	26	28	21	25	31	25	17	19	37	19	08	28	37	33	....	39	28	17	22	21	36
25	53	17	39	22	17	42	29	30	19	16	31	18	34	16	30	08	09	07	18	11	16	24	36	21	41	14	21	27	00	15	40	32	29	....	17	-02	08	48	36
18	33	26	20	18	22	26	25	27	24	30	21	22	32	35	37	24	22	23	18	25	17	27	21	17	05	15	32	13	03	45	36	18	28	17	....	15	17	05	23
17	12	38	22	23	31	11	25	29	14	33	12	27	08	12	25	27	26	16	31	34	23	19	02	38	-13	28	52	-13	14	23	10	29	17	-02	15	....	67	-13	02
21	20	28	29	22	32	16	25	29	15	37	17	30	16	13	28	26	27	15	37	35	28	20	06	41	-07	32	55	-11	12	23	17	35	22	03	17	67	....	-05	05
14	29	-01	12	07	02	20	09	09	08	01	17	02	13	07	13	-04	00	-03	04	01	01	09	18	04	60	06	02	35	-07	01	18	18	21	48	06	-13	-05	....	45
14	26	09	11	15	10	33	14	09	22	05	15	08	18	10	17	02	08	13	08	09	02	06	24	02	31	11	14	37	05	15	23	27	39	38	23	02	05	45	....
25	27	14	26	12	11	26	16	14	16	22	21	26	24	10	23	16	17	13	27	17	22	19	22	26	11	31	27	13	07	14	28	41	26	31	14	13	20	25	29
35	39	23	40	21	22	28	26	28	16	40	27	38	29	13	31	20	23	17	37	31	33	28	17	46	05	40	49	00	14	16	30	52	30	35	12	40	50	13	14
11	20	15	13	13	17	23	15	18	16	03	10	05	15	11	18	14	12	05	12	03	07	11	19	08	05	-02	14	13	05	16	13	04	21	22	15	03	04	05	15
04	20	13	12	09	12	26	[illegible]	07	10	-01	14	01	10	05	12	00	02	00	08	02	-01	06	20	02	19	-09	05	27	00	05	14	-02	22	35	05	-07	-05	25	25
15	23	22	20	16	23	17	[illegible]	27	13	14	05	12	13	18	22	20	15	11	14	17	15	17	13	21	10	08	23	08	08	19	13	10	15	08	29	15	14	[illegible]	[illegible]
23	35	28	25	24	32	40	[illegible]	33	26	21	30	23	28	25	39	25	25	11	24	19	21	28	26	23	17	03	36	18	07	30	29	20	35	32	28	16	12	18	22
15	24	18	20	19	20	23	[illegible]	24	12	18	24	15	19	12	[illegible]	17	18	05	15	[illegible]	15	20	18	14	[illegible]	[illegible]	22	21	[illegible]	16	19	[illegible]	25	34	14	[illegible]	[illegible]	[illegible]	[illegible]
31	35	32	42	24	32	39	[illegible]	38	25	36	29	34	34	22	30	34	34	22	38	31	25	29	21	30	04	22	45	10	16	30	31	23	25	27	26	26	30	01	[illegible]

34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	M	σ	Test No.
24	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	25	34	31	25	18	26	19	[illegible]	[illegible]	26	25	25	07	01	[illegible]	[illegible]	03	27	[illegible]	1.93	1.
19	24	31	45	36	34	41	22	46	[illegible]	31	40	08	10	30	42	49	31	35	26	25	34	15	13	34	50	05	09	16	31	17	37	4.69	1.97	2.
18	20	27	39	30	33	42	35	39	21	37	49	15	06	26	48	45	34	48	27	19	26	25	21	35	52	11	14	17	30	19	36	4.70	2.01	3.
08	10	07	27	09	23	21	18	25	05	15	24	09	01	13	27	23	22	27	14	07	13	14	12	20	26	10	10	10	20	10	20	4.78	1.94	4.
17	23	24	36	31	15	24	19	34	06	79	35	−04	07	21	27	55	24	12	21	33	34	05	27	33	35	06	−04	08	13	−01	27	4.94	1.99	5.
24	23	20	17	24	14	29	−01	26	05	10	36	06	07	37	21	17	23	08	32	26	26	02	09	06	19	15	03	19	22	14	23	4.60	1.98	6.
24	27	19	27	23	19	30	−02	25	00	12	33	05	03	32	18	14	20	05	28	26	26	03	12	03	18	11	02	22	20	15	19	5.09	1.92	7.
32	38	26	35	43	20	35	12	39	07	18	51	03	09	47	24	25	21	02	42	39	38	−03	02	12	24	14	01	27	32	13	33	4.71	1.93	8.
18	17	34	17	06	16	28	21	19	21	07	26	14	05	29	39	25	33	28	37	04	08	16	38	15	15	18	11	11	28	19	29	4.93	2.16	9.
32	30	14	30	14	21	39	28	40	13	12	46	08	06	34	41	32	36	37	31	21	26	15	19	23	38	14	14	12	32	17	35	4.91	2.08	10.
23	27	14	35	11	20	43	42	36	28	24	37	16	08	31	47	40	45	50	30	14	13	25	37	25	34	19	19	14	36	22	41	4.94	1.96	11.
19	30	15	33	−07	21	30	22	36	25	−02	36	19	06	32	40	25	31	36	26	08	14	24	22	22	26	18	16	16	36	25	32	5.06	1.91	12.
27	25	18	18	08	14	33	19	25	14	02	30	09	05	23	30	16	22	29	21	12	13	11	10	11	22	21	18	11	33	20	24	5.12	1.50	13.
22	23	09	37	00	31	36	21	39	14	06	38	09	06	33	33	21	25	29	23	13	16	10	11	20	26	12	14	14	28	17	31	4.61	2.22	14.
20	25	13	33	15	24	36	34	38	25	16	36	16	04	28	48	35	43	49	29	10	17	20	24	26	35	14	17	14	34	21	37	4.93	1.74	15.
11	12	17	13	00	19	29	12	17	05	05	22	03	06	19	10	04	10	13	12	15	18	−01	−02	04	15	12	07	14	12	10	14	4.92	2.00	16.
16	21	22	22	15	19	26	19	23	04	18	27	11	07	30	29	21	22	15	28	13	13	05	10	13	18	16	05	17	20	08	17	4.86	1.92	17.
14	11	23	30	25	25	26	25	33	16	41	31	07	08	20	27	38	27	25	18	17	21	14	14	25	35	11	04	15	23	15	31	4.57	1.96	18.
15	19	32	28	28	31	49	44	37	35	28	38	22	03	32	52	34	43	53	33	12	20	29	26	27	39	20	20	23	35	24	35	4.83	1.93	19.
30	31	18	35	38	34	28	08	43	−02	13	40	05	10	30	25	16	23	17	28	23	23	−01	09	14	23	15	13	22	28	18	32	4.81	1.94	20.
19	20	28	34	38	23	33	31	37	17	38	42	12	10	26	23	41	33	39	20	22	29	12	11	26	40	13	12	20	29	20	42	4.77	1.92	21.
26	25	20	23	27	16	37	19	34	07	17	34	11	12	25	20	18	24	22	18	23	22	07	15	12	21	13	09	16	24	19	24	5.41	2.00	22.
16	23	17	20	44	18	33	15	30	09	07	41	03	13	32	20	15	30	17	22	31	32	02	10	11	22	17	12	23	32	20	32	5.00	1.90	23.
16	23	18	29	31	27	26	39	33	12	23	35	21	09	28	42	34	51	42	26	11	16	20	33	26	28	23	26	17	40	23	39	4.73	1.92	24.
20	27	20	27	33	29	39	20	39	15	24	36	06	11	28	25	23	31	29	25	25	25	09	14	16	26	15	09	30	33	19	35	5.05	1.93	25.
25	23	11	32	35	28	35	18	40	13	21	40	07	09	31	28	24	28	30	27	29	29	09	09	14	28	18	07	27	33	24	35	5.01	2.03	26.
24	21	40	17	30	13	19	23	15	10	19	28	09	03	25	24	28	27	19	24	14	15	08	22	16	16	16	10	13	26	12	25	5.09	1.93	27.
31	33	20	32	32	27	31	16	37	01	31	40	01	14	37	33	38	27	15	30	33	37	01	05	22	40	03	−01	14	21	18	36	5.13	1.92	28.
27	27	10	30	27	27	32	31	30	15	24	33	15	11	30	33	25	25	31	21	12	17	17	15	21	27	10	14	03	30	24	29	5.03	1.96	29.
29	33	23	44	39	25	34	15	51	02	45	44	03	13	27	29	41	23	18	22	27	30	02	08	26	38	05	01	12	23	15	34	4.82	1.97	30.
20	20	21	20	27	19	26	39	24	16	24	33	16	01	39	44	27	34	34	32	08	16	13	18	24	29	15	10	13	28	19	34	4.90	1.94	31.
19	17	13	13	19	11	30	22	17	13	05	23	11	06	32	25	13	16	16	35	12	13	07	10	10	13	11	05	18	25	12	22	5.00	1.69	32.
31	35	17	36	37	26	40	31	44	18	24	47	11	08	44	33	25	35	30	37	25	23	13	17	23	31	18	12	22	39	25	39	4.91	1.84	33.
	48	28	21	32	21	30	14	30	00	24	39	03	14	31	19	25	19	03	24	27	26	−04	02	16	20	14	00	20	25	17	34	4.92	1.90	34.
45		15	35	34	26	25	12	32	02	26	41	04	14	30	19	25	21	09	22	26	27	00	08	17	23	12	02	15	25	18	34	4.86	1.93	35.
28	15		17	26	13	25	15	19	−01	24	27	−04	11	20	22	25	21	07	23	16	15	−03	13	13	17	05	00	11	11	05	22	5.33	2.06	36.
21	35	17		30	29	29	12	42	−01	23	45	−01	07	23	25	35	26	18	13	31	37	04	06	27	37	[illegible]	08	14	24	15	33	5.06	1.98	37.
32	34	26	30		11	17	07	33	−01	25	41	−02	08	22	24	27	28	11	25	34	35	01	08	17	31	08	−02	17	19	09	31	4.97	2.01	38.
21	26	13	29	11		23	16	29	04	18	31	−08	09	17	23	24	21	16	17	23	28	01	02	22	33	07	−01	15	21	15	25	5.19	1.93	39.
30	25	25	29	17	23		24	36	16	26	35	01	[illegible]	34	33	27	25	24	27	19	20	09	06	19	28	11	06	17	26	20	29	5.18	1.93	40.
14	12	15	12	07	16	24		15	19	14	19	15	−01	20	35	24	31	36	21	02	06	18	24	22	17	19	20	13	26	18	21	5.06	1.97	41.
30	32	19	42	33	29	36	15		04	28	52	−04	06	27	29	41	25	21	17	38	41	04	02	26	46	08	02	21	23	14	39	4.94	1.84	42.
00	02	−01	−01	−01	04	16	[illegible]	04		01	02	32	−09	04	19	10	17	41	06	−13	−07	60	31	11	05	06	19	10	17	22	04	5.10	2.01	43.
24	26	24	28	25	18	26	14	28	01		32	−14	08	13	26	57	19	14	15	28	32	06	11	31	40	−02	−09	08	09	03	22	5.03	1.94	44.
39	41	27	45	41	31	35	19	52	02	32		03	17	40	35	41	37	21	32	52	55	02	14	27	49	14	05	23	36	22	45	5.68	2.12	45.
03	04	04	−01	−02	−03	01	15	−04	32	−14	03		20	08	13	01	19	27	13	−13	−11	35	37	13	00	13	27	08	18	21	10	5.10	1.97	46.
14	14	11	07	03	09	04	−01	06	−09	08	17	20		09	04	13	03	00	08	14	12	−07	06	07	14	05	00	08	07	02	16	5.06	2.01	47.
31	30	20	23	22	17	34	20	27	04	13	40	08	09		39	18	23	15	45	23	23	01	15	14	16	15	05	19	30	16	30	5.03	2.00	48.
19	19	22	23	24	23	33	35	29	19	26	35	13	04	39		33	37	40	36	10	17	18	23	28	30	13	14	13	29	19	31	4.98	1.88	49.
25	23	25	35	27	24	27	24	41	10	57	41	01	13	18	33		33	32	18	29	35	18	27	41	52	04	02	10	20	09	33	5.35	2.03	50.
19	21	22	28	23	21	25	31	25	17	19	37	19	08	28	37	33		39	23	17	22	21	36	26	30	20	22	15	36	25	36	5.08	1.97	51.
08	09	07	18	11	16	24	36	21	41	14	21	27	00	15	40	32	39		17	−02	08	48	36	34	35	22	35	03	32	34	27	5.22	1.97	52.
24	22	23	18	25	17	27	21	17	05	15	32	13	08	45	36	18	28	17		15	17	06	23	14	12	15	05	20	28	14	28	5.37	1.93	53.
27	26	16	31	34	23	19	02	33	−13	28	52	−13	14	23	10	29	17	−02	15		67	−13	02	13	40	03	−07	15	16	04	26	4.98	2.04	54.
26	27	15	37	35	28	20	08	41	−07	32	55	−11	12	23	17	35	22	03	17	67		−08	05	20	50	04	−05	14	12	06	30	5.12	2.10	55.
−04	00	−03	04	01	01	03	18	04	60	06	02	35	−07	01	18	18	21	48	06	−13	−08		45	25	13	08	28	03	18	28	04	5.14	1.98	56.
02	08	13	06	08	02	06	24	02	31	11	11	37	05	15	23	27	35	36	23	02	05	45		29	14	15	29	08	22	19	16	5.20	1.97	57.
16	17	13	27	17	22	19	22	26	11	31	27	13	07	14	28	41	26	34	14	13	20	25	29		34	09	13	08	20	14	24	5.17	1.96	58.
20	23	17	37	31	33	28	17	46	05	40	40	00	14	16	30	52	30	35	12	40	50	13	14	34		01	02	09	18	11	31	5.41	2.03	59.
14	12	05	12	05	07	11	19	08	05	−02	14	13	05	16	13	04	20	22	15	03	04	08	16	03	01		35	30	39	33	23	4.63	1.91	60.
00	02	00	08	02	−01	06	20	02	19	−09	05	27	00	05	14	−02	22	35	05	−07	−05	28	29	13	02	35		11	30	40	15	4.88	1.93	61.
20	15	11	14	17	15	17	13	21	10	08	23	08	08	19	13	10	15	03	20	15	14	03	03	03	00	30	11		29	23	27	4.70	1.89	62.
25	25	11	24	19	21	26	26	23	17	09	36	18	07	30	29	20	36	32	23	16	12	18	22	20	18	39	39	29		49	43	5.40	1.94	63.
17	18	05	15	09	15	20	18	14	22	03	22	21	02	16	19	09	25	34	14	04	06	28	19	14	11	33	40	23	49		27	4.84	1.96	64.
34	31	22	34	31	25	29	21	39	04	22	45	10	16	30	31	33	36	27	26	26	30	04	16	24	31	23	15	27	43	27		5.21	1.96	65.

APPENDIX D

Intercorrelations, Means, and Standard Deviations of Tests in the Air-Crew Classification Battery of September, 1944[1]

[1] The examinees on which these data are based consisted of 8,158 unclassified aviation students tested at Sheppard Field, Tex. The scoring formulas used are presented in report no. 2 in the series of AAF aviation psychology research reports. *Vid.* DuBois, P.M. *op. cit.* Table 3.18.
Decimal points that properly precede entries in the matrix have been omitted.

TEST NUMBER

TEST NO.	TEST TITLE	CODE NO.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	M	σ
1.	Dial & Table Reading	CP621-622A		028	166	400	346	414	371	212	317	516	548	022	145	267	492	138	044	226	365	217	449	5.682	2.122
2.	Biographical Data (P)	CE602D	028		198	079	131	011	192	272	127	−132	−112	350	366	132	001	133	274	077	182	211	095	5.100	1.968
3.	Biographical Data (N)	CE602D	166	198		092	044	126	094	005	078	139	122	−074	064	072	135	051	093	090	074	017	155	5.065	2.010
4.	Spatial Orientation I	CP501B	400	079	092	....	388	177	280	151	454	233	232	006	148	139	163	164	054	193	301	165	302	5.034	2.002
5.	Spatial Orientation II	CP503B	346	131	044	388	..	327	372	399	361	104	172	182	228	275	299	131	136	132	292	187	311	4.983	1.870
6.	Reading Comprehension	CI614H	414	011	126	177	327	....	333	325	176	299	351	179	271	407	516	043	021	096	202	093	333	5.353	2.028
7.	Instrument Comprehension	CI616C	371	192	094	280	372	333	....	393	279	168	218	214	365	256	295	201	220	152	355	247	359	5.080	1.973
8.	Mechanical Principles	CI902?	212	272	005	151	399	325	393	.....	171	−015	052	483	357	340	351	216	347	076	320	340	272	5.224	1.966
9.	Speed of Identification	CP610A	317	127	078	454	361	176	279	171	...	150	174	062	231	135	122	149	053	190	278	142	263	5.371	1.970
10.	Numerical Operations I	CI701B	516	−132	139	233	104	299	168	−015	150	...	668	−131	020	132	404	026	−074	153	158	043	265	4.980	2.043
11.	Numerical Operations II	CI701B	548	−112	122	232	172	351	218	052	174	668		−062	052	200	304	037	−032	145	124	065	303	5.119	2.102
12.	Mechanical Information	CI905B	022	350	−074	006	183	179	214	483	062	−131	−062		446	247	130	076	278	032	178	283	042	5.140	1.960
13.	General Information	CE505F	145	366	064	148	229	271	365	337	231	020	052	446		288	138	164	289	079	221	194	159	5.203	1.972
14.	Practical Judgment	CI301C	267	132	072	139	275	407	256	340	135	132	200	247	288	..	343	086	126	083	199	140	242	5.174	1.965
15.	Arithmetic Reasoning	CI206C	492	001	135	163	299	516	295	351	122	404	304	130	138	343		008	018	089	185	108	314	5.412	2.026
16.	Rotary Pursuit	CP410B	138	133	051	164	131	042	201	216	148	026	037	076	164	086	008		351	296	380	333	229	4.629	1.911
17.	Rudder Control	CM120B	048	274	003	054	136	021	220	347	053	−074	−052	278	289	126	018	351	•	111	394	400	152	4.860	1.920
18.	Finger Dexterity	CM116A	226	077	090	193	132	096	152	076	199	153	143	032	079	083	096	206	111	...	292	229	268	4.696	1.585
19.	Complex Coordination	CM701A	365	182	074	301	292	202	358	320	276	158	124	178	221	199	155	359	394	292	.	492	430	5.401	1.936
20.	Two-Hand Coordination	CM101A	217	211	017	165	187	093	247	340	142	043	065	283	194	140	108	333	400	228	492	...	269	4.838	1.960
21.	Discrimination Reaction Time	CP611D	449	095	155	302	211	333	358	272	263	265	305	042	159	212	314	229	152	268	430	269		5.209	1.962

Index of Names

703320—47—59

Subject Index

C

G

J

K

L

M

☆ U. S. GOVERNMENT PRINTING OFFICE: 1947—[illegible]