AFCAPS-FR-2019-001C

Army Air Forces Aviation Psychology Program Research Reports

Problems and Techniques

Report No. 3

1947

Edited by

Robert L. Thorndike

Air Force Personnel Center Strategic Research and Assessment HQ
AFPC/DSYX
550 C Street West, Ste 45 Randolph AFB TX 78150-4747

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report was cleared for release by HQ AFPC/DSYX Strategic Research and Assessment Branch (SRAB) and is releasable to the Defense Technical Information Center.

This report is published as received with minor grammatical corrections. The views expressed are those of the authors and not necessarily those of the United States Government, the United States Department of Defense, or the United States Air Force. In the interest of expediting publication of impartial statistical analysis of Air Force tests SRAB does not edit nor revise Contractor assessments appropriate to the private sector which do not apply within military context.

Federal Government agencies and their contractors registered with Defense Technical Information Center should direct request for copies of this report to:

Defense Technical Information Center - http://www.dtic.mil/

Approved for public release, unlimited distribution by AFPC/DSYX Strategic Research and Assessment Branch (SRAB) Joint Base San Antonio-Randolph AFB, TX 78150-4747 or higher DoD authority. Please contact AFPC/DSYX Strategic Research and Assessment Branch (SRB) with any questions or concerns with the report.

This paper has been reviewed by the Air Force Center for Applied Personnel Studies (AFCAPS) and is approved for publication. AFCAPS members include: Senior Editor Dr. Thomas Carretta AFMC 711 HPW/RHCI and Dr. Imelda Aguilar HQ AFPC/DSYX.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
1947	Final Report	

4. TITLE AND SUBTITLE

Army Air Forces Aviation Psychology Program Research Reports: Research Problems and Techniques, Report No. 3

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Edited by Robert L. Thorndike

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
AFPC/DSYX
Strategic Research and Assessment Branch
Randolph AFB TX, 78150

8. PERFORMING ORGANIZATION REPORT

AFCAPS-FR-2019-001C

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Personnel Center
Strategic Research and Assessment Branch
Randolph AFB TX, 78150

10. SPONSOR/MONITOR'S ACRONYM(S)
HQ AFPC/DSYX

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFCAPS-FR-2019-001C

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release. Distribution Unlimited

13. SUPPLEMENTARY NOTES
Task 3: Review of Historical Aviation Constructs; USAF Strategic Personnel Research Program(DTIC Access Number: AD1078664)

During World War II, the Army Air Force Aviation Psychology Program (AAP) conducted extensive research on selecting and training aircrew. The results of this program were documented in a 19-volume series of research reports. Damos Aviation Services, Inc. (DAS), prepared the series for re-publication in a digital format. Report No. 3 summarizes the procedures which were developed and the problems which were encountered in the Aviation Psychology Program of the Army Air Forces; in particular, problems concerning the selection and classification of personnel for air crew assignment.

15. SUBJECT TERMS
Aviation Psychology, Blue Books, Digitize, Index, Personnel Selection and Classification, Air Crew Assignment

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
Unclassified					Katie Gunther, Ph.D.
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U	U	175	210-565-5245

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Army Air Forces
Aviation Psychology Program
Research Reports

Research Problems and Techniques

REPORT NO. 3

Edited by
ROBERT L. THORNDIKE
Associate Professor of Education
Teachers College
Columbia University

1947

For sale by the Superintendent of Documents, U. S. Government Printing Office
Washington 25, D. C. - Price $1.00 (Buckram)

Preface

The present report undertakes to summarize the procedures which were developed and the problems which were encountered in the Aviation Psychology Program of the Army Air Forces, and more particularly in that portion of it which was concerned with the selection and classification of personnel for aircrew assignment. An attempt has been made to record the procedures and problems which seemed of general and continuing interest, so that the techniques and insights achieved in the war just past may be available to future students and workers in the field.

In this report, no attempt has been made to prepare a complete statement of statistical methodology. For supplementary material on both the theory and computing routines for statistical analysis, the reader is referred to Report No. 18, Records and Analysis Procedures. Additional statistical material, with special detail on the procedures and results of factor analysis, will be found in Report No. 5, Printed Tests. Still further statistical and methodological material will be found in Report No. 6, The AAF Qualifying Examination. The content of this report probably slights research on training procedures and on proficiency measurement. This arises in part because the problems were less unified and homogeneous, in part because less methodological contribution was made in these fields, in part probably because the editor was less well acquainted with this part of the research program.

The procedures, techniques, and insights which are reported here were the joint achievement of a great many people scattered throughout the Aviation Psychology Program. Acknowledgement has been made in footnotes of the authorship of certain specific formulas or of responsibility for certain specific studies, but there is much more which does not lend itself to such acknowledgement. The ideas of many workers in the program, and particularly of its Director, Col. J. C. Flanagan, appear on every page. The editor has been responsible only for assembling, organizing, and interpreting these ideas. He can claim only very limited personal contribution to them, but he must take full responsibility for any inadequacy in reporting or interpreting.

Except for a few brief sections written by Lt. Col. A. P. Horst, the report was written by the editor. The manuscript has been read in part by Cols. J. C. Flanagan and J. P. Guilford and by Lt. Cols. A. P. Horst and M. P. Crawford, and has profited from their criticisms and suggestions.

ROBERT L. THORNDIKE, *Major, A. C.*

WASHINGTON, D. C., MAY 1946.

Contents

CHAPTER ONE

General Introduction

It is the purpose of this volume to present a general discussion of the research problems which were encountered in the Aviation Psychology Program in the Army Air Forces and of the methods and procedures which were developed for dealing with those problems. Succeeding reports will present in detail the results of the activities of the various organizations working within the program. The reports will describe and evaluate the various types of test materials which were developed for aptitude testing and will cover the general research activities in connection with the improvement of proficiency measures and with studies of training procedures for each of the aircrew specialties. In the present report specific data will be presented only insofar as they are needed to provide illustrations of the problems which were encountered and the procedures which were developed. No attempt will be made systematically to cover the data for their own sake.

The psychological research program in the Army Air Forces may be divided into two major phases. The first of these to be undertaken was the development of testing procedures for use in the original selection and classification of personnel for assignment to the various aircrew specialties, with particular attention to the specialties of pilot, navigator, and bombardier. This was quite a unified program, with a rather well defined and precise objective. The objective can be stated as the development of procedures for the assignment of personnel to one of a number of training specialties which would maximize the effectiveness of subsequent training and combat operations. We shall see that the objective becomes somewhat less well defined upon detailed analysis, as we try to deal with it in terms of specific operations for classification. However, relatively speaking, the research problem remained a homogeneous and unified one.

In the second phase, the research program branched out from the initial research in original selection and classification to the study of all types of psychological problems relating to the ultimate effectiveness of combat personnel. In addition to initial selec-

tion and classification, attention was devoted to the improvement of training methods, objective methods for evaluating proficiency, later classification into special duty assignments, evaluation of combat leadership, equipment design, and a variety of other problems. Though still held together by the general theme of maximizing combat effectiveness, these studies were not unified to the same degree as the earlier work in selection and classification.

The lack of unity in the later work makes a systematic treatment of research problems in those areas difficult. For that reason, this report has been organized basically around the problems of selection and classification. Many of the same problems, such as those of criteria of proficiency or those of determining reliability, enter into the various other types of research projects which were subsequently undertaken. In fact, the number of entirely novel problems introduced by training research and the like, as opposed to selection and classification research, is not believed to be great. An attempt has been made to discuss a few of the problems which were unique to training research in the last chapter.

The sequence of chapters for this report follows in a general way the sequence of operations in test development. Chapter headings are as follows:

CHAPTER TWO

Job Analysis Problems and Procedures

GENERAL INTRODUCTION

As indicated in Chapter 1, the first phase of the research program in the AAF was the development of procedures for selecting men for different aircrew specialties. As in any program of personnel selection for a certain number of specialized jobs, the first step, logically and to a certain extent chronologically, was an analysis of the jobs in question to determine the activities which were carried out in those jobs, the circumstances under which they were carried out, and the psychological traits or functions which appeared to be important in carrying out those activities.

When the psychological research program was first established under the jurisdiction of the Air Surgeon to do research on the selection of men for pilot training, it immediately became obvious that the first need was for better information as to the characteristics of the job of the pilot. With each subsequent expansion of the scope of the program to include, first, bombardier and navigator selection and, later, selection for flight engineer, radar operator, gunner, and various types of specialized enlisted aircrew, the necessity for analysis of the new job specialties continued to be evident. At the very beginning of the research program, then, and continuing throughout the program a good deal of research effort was devoted to the problems of job analysis. Major job analysis studies were assembled in a series of Analysis of Duties Bulletins which were distributed as they were issued to all officers and units concerned with the research program. It will be appropriate at this time to consider the various approaches which were made to studying the different aircrew jobs and to attempt an evaluation of the contribution and of the limitations of each of these approaches.

REVIEW OF LITERATURE

Naturally the first source to which a scientist turns in dealing with any problem is the existing literature of the topic. The studies

which other men have carried out and the generalizations which have resulted from their studies provide the initial orientation in terms of which further work is done. The value of this approach varies with the specific case, depending upon the quantity and quality of previous work. The amount of directly relevant material on most aircrew specialties was rather limited. However, the material is already organized and available for study, and provides the natural jumping-off place for future work. The information which the newcomer to a field can extract from these materials is always limited by the limits of his own background. Purely verbal presentations cannot supply a concrete background of experience for understanding any job. Insofar, however, as the literature has been prepared by professional psychologists and well formulated in psychological categories, it promises more of value than many of the non-professional records.

ANALYSES OF RECORDED MATERIALS

One approach to analysis of aircrew specialties was through materials which were available in written form. The written materials fell into two broad categories:

1. Technical manuals, curricula, and other general instructional materials developed for the purpose of defining the program of instruction or providing instructional aids.

2. Records providing comments upon the performance of specific individuals in training or of groups in combat duty.

The first of these types of material is pretty much self-explanatory. It is obvious that in order to conduct a large-scale program of flying training it is necessary to have guides for controlling and standardizing the material which is taught, together with materials to supplement actual flying instruction. An examination of these materials provides the psychologist with a general orientation as to the nature of the task involved. He can get a general picture of what the individual is expected to learn during each stage of training for the aircrew specialty under study. From the point of view of the psychologists developing the aircrew testing program, the outstanding advantage of these materials was their availability. They could ordinarily be obtained at any headquarters or training station. The materials were studied with varying degrees of thoroughness by many individuals concerned with developing the testing program. The limitations of these materials are fairly obvious. They were, after all, purely verbal presentations and removed from the realities of the actual task. They dealt typically with fairly gross units of behavior to be learned and results to be achieved rather than with very detailed reports of activities to be carried out or of conditions under which they were to be carried out. They evaluated results to be attained

rather than psychological traits important for attaining them. They provided, therefore, only an indirect and rather remote set of cues to the actual psychological functions for which tests were desired.

The second type of record merits somewhat more detailed consideration. In the aircrew training program a great number of different types of records of performance were maintained for individuals in training. Most of these records were of interest in aptitude test development only because of their possible use as criteria. In most cases they were in the form of quantitative grades or ratings rather than in the form of qualitative descriptions. There were, however, a certain number of qualitative and descriptive records maintained in connection with the program of training, and these presented some possibilities of providing information concerning traits important for effective performance of the job. At each level of pilot training, for example, daily grade slips were made out for each man. These grade slips contained not only quantitative grades on various maneuvers and phases of flying but also comments on the nature of the student's performance in any maneuver in which he was judged to be deficient.

Another type of record from which some clues as to the problems and difficulties encountered by the cadet during training were obtained were the records of Elimination Board proceedings available for each cadet who was eliminated from flying training. Here testimony by the instructor was available as to the particular deficiencies of the cadet in question. Analysis of a number of these Board proceedings brought out the recurring patterns in instructor comments and provided a basis for setting up certain categories with regard to reasons for elimination. The same type of material at a more advanced stage is seen in the Reclassification Board proceedings which were held for men who had received their commission as fliers but whose subsequent performance had been so unsatisfactory as to make it seem necessary to reconsider their flying status.

A further type of record found in a few cases, and of particular interest because of it direct relevancy to the task of personnel in combat operations was the analysis of reasons for failure of combat missions. Mission reports were available for certain groups in combat theaters, and had in certain cases been assembled and analyzed so as to indicate what had gone wrong on a number of unsuccessful missions. These records pointed out certain recurring deficiencies of combat personnel, and gave information as to factors which needed to be taken into account in either selection or training of personnel, or both, if the efficiency of the combat team was to be improved.

Materials of the sort which have just been described appear to get somewhat nearer the determination of actual psychological functions involved in aircrew performance than do the general descriptions of curriculum and training procedures. It was possible to classify most of the comments on grade slips or in Board Proceedings into categories according to the psychological functions which appeared to be involved. Thus certain remarks were classified as indicating deficiencies of memory, in other cases cadets were said to have shown poor judgment and the like.

In the evaluation of materials such as Elimination Board proceedings, certain features of the record made them seem quite promising. In the first place, the record sumarized impressions based upon a good deal of rather intimate experience with the man in question, since it was the final summary evaluation based upon all his training at the station in question. The evaluation combined the judgments of several men, instructors and check riders, who had a varied and often wide experience of instructing cadets. It may be contended that the seriousness of the use to which the results were put is an evidence that evaluations were carefully and deliberately made. The statements were subject to rebuttal by the cadet at the hearing so that they had to seem reasonable and appropriate to him. In other words, the evaluations represented a practically important and therefore carefully and conscientiously rendered evaluation of individual ability.

The materials were felt to present certain difficulties however, both as to the adequacy of the original records and as to the interpretation of the reports in categories useful for test construction. First of all, in view of the number of reports of this type which had to be made we must expect to find a tendency for the making of them to have become rather perfunctory and stereotyped. It seems clear that instructors who were concerned with the evaluations developed a certain stock set of categories which they applied somewhat uncritically to new individuals as they came along. As a matter of fact, in the case of the grade slips, the development of such stock phrases was even encouraged by the publication of a standard set of comments from which the instructor was invited to select the one to be applied to the deficiency of a particular cadet upon a particular unit of instruction. Insofar as the set of standard remarks was inclusive and the individual comments were well phrased, this may well have improved the quality of instructor criticism. At the same time, however, it would have tended to limit the spontaneous range of remarks and may possibly have led to the exclusion of certain significant categories. The job analysis is then based upon a second-order abstraction of a set of categories from a set of remarks which had already been abstracted from concrete experience.

Another difficulty was felt to lie in a tendency to make a strong case in connection with the elimination or reclassification of any individual. If a person was being recommended for elimination it was only natural to try to make the report on the individual look as clear-cut and decisive as possible. Stock comments might be used not because they were particularly appropriate to an individual currently being considered for elimination but because they were part of the accepted pattern of reasons offered for elimination.

Finally, there appeared to be a certain amount of difficulty with language. Terms were used with meanings which varied from one report to another, and meanings which were perhaps at variance with the meaning of the same term to the psychologist working on the problem. This can be well illustrated by the term "judgment." Poor judgment was repeatedly offered as a reason for failure in flying training, either upon a single maneuver or for the whole course of training. Further inquiry into the exact meaning of this term "judgment" revealed that it meant various things at various times and places. At one time and to one person it meant lack of common sense represented by a decision to fly through a storm rather than returning to the starting place. At another time it meant a bad decision in the quick choice of an emergency landing field for a simulated forced landing. At still another time it meant inaccurate perception of the relative speed and position of the cadet's plane and another plane. In still other cases it meant variations of these and other types of judgments, intellectual or perceptual, which the individual was called upon to make. On this basis it is not difficult to see that reports of cadet failure because of poor judgment were only moderately instructive to the psychologist trying to determine the specific functions for which tests should be constructed. Language is sufficiently a source of confusion in communication between trained psychological personnel; it became even more so in working with flying personnel who were not chosen on the basis of verbal facility and who had not been trained to be precise or analytical in their reports of human behavior.

INTERVIEWS WITH AND INTERROGATIONS OF PERSONNEL

There were a number of projects, during the war, in which personnel concerned with or involved in flying training were interviewed by aviation psychologists to obtain their reactions to various features of their experience. A great deal of this form of interaction took place in informal personal contacts. In certain instances it was also formalized into definite interview projects. In general, three types of personnel were interviewed:

persons presumably proficient and experienced in the job under study, persons undergoing training, and persons having failed in the task.

Interviews with persons presumably proficient and experienced in the particular job tended to cover two types of groups. In the first place, as a part of general orientation to the duties and problems of personnel in training, interviews were carried out with instructors, directors of training, and other personnel in the training program who had a chance to observe at first hand the course of training operations. In the second place, interviews were carried out with personnel returned from combat theaters. These interviews had as their particular purpose the description of flying activities carried on under combat conditions and the determination of the distinctive qualities necessary for success in combat operations as distinct from those necessary for successful completion of training. An approach of particular interest in this connection was the systematic interrogation of returned combat personnel on the causes of mission failures. In this work, testimony was obtained from returnees at Redistribution Stations as to the mission in which they had participated which had been least successful. Data were obtained as to the nature of the mission and the type of personnel failure contributing to the poor mission.

Personnel undergoing training were interviewed with a view to determining the particular problems and difficulties which they were currently experiencing in the course of their training activities. For example, evidence of various sorts had consistently pointed to landing as one of the crucial tasks at an early stage in pilot training. One procedure for learning more about the particular nature of landing difficulties was to interview a number of cadets in primary training who were passing through the stage of learning to land.

Interviews with persons failing the task were carried out in order to point out still further difficulties encountered in the task: A study was carried out interviewing eliminees shortly after the time of their elimination from pilot training, for example, to determine the problems and difficulties which had been experienced by this particular group of students for whom problems and difficulties had presumably been particularly acute.

There was a definite benefit to be derived from interviews with persons having serious difficulty in mastering essential features of the training program as contrasted with experts in the specialty or students who were having relatively little difficulty. The recognized experts, such as instructors, directors of training, and combat returnees, did not have first-hand knowledge of difficulties and could only speak, therefore, from observation of those who

did; whereas those who learned with less facility were in a position to point out in some detail those parts of the training curriculum which caused them most trouble and to indicate why, in their opinion, they had been unable to master specific problems. On the other hand a certain amount of caution is necessary in interpreting information obtained from unsuccessful trainees. Frequently there is a tendency to offer rationalizations which assign the responsibility of their failures to others than themselves.

The interview procedures which we have just discussed have as advantages over the procedures previously described the fact that they are more flexible and permit follow-up in more detail along those leads which appear novel or promising. They provide a way of exploring any areas which may be suggested, either by the interviewer or by the person interviewed. The difficulty of communication is still maintained, in that the psychologist is at least one step removed from the actual situation and can experience it only as it is reported to him by the person interviewed. This difficulty is aggravated by the fact that flying personnel are typically not highly articulate about their own experiences. It is probably safe to say that the flier is typically more a man of deeds than of words. Job analyses which proceed through the technique of asking the flier to indicate the qualities necessary for success in the task in which he has engaged in a sense substitute the flier's untrained analysis of the functions which enter into flying success for the trained analysis of the professional psychologist. Again, a question may be raised as to the amount of insight which individuals will have into reasons for their own success or failure in training when they have had no special background or special motivation to develop this insight.

DIRECT PERSONAL EXPERIENCE BY PSYCHOLOGICAL PERSONNEL

When the program for developing a battery of aircrew classification tests was initially undertaken, there was no time for psychological personnel to receive training in the types of jobs for which selection would have to be carried out. The initial development of tests for a battery had to proceed very promptly and it was necessary that the most expeditious job analysis procedures be relied upon to provide guides for test construction. However, as the testing program continued it was more possible for psychological personnel to experience at first-hand at least the initial levels of training for the various types of aircrew duty.

The first direct experience of psychological personnel with the aircrew specialties for which testing was being carried out took the form of relatively brief visits to and inspections of training

703315—47—2

stations. These were possible almost at once and in these visits psychological personnel were given opportunity to participate in sample training missions, visit classes, interview supervisory and instructional personnel, examine at first-hand the types of instruments used by men engaging in the various aircrew specialties, and see the training program in actual operation.

These brief visits seemed to lend concreteness to the concepts of the activities carried out by personnel being trained in the various aircrew specialties and of the conditions under which they were carried out. They provided an over-all view of the training situation and aided in arriving at job descriptions and analyses both through the insights obtained upon the visit and through the better background which the visits gave for understanding and interpreting other more indirect sources of information. Obviously, they provided only a limited amount of direct *experience* of the job in contrast to *observation* of it, due to limitations of time. The psychologist had opportunity to observe a good deal of what was done and learned, but to learn relatively little of it himself.

As the research program continued, time was finally made available for certain aviation psychologists to take substantial amounts of training in the basic stages of several of the aircrew specialties. Two officers went through primary pilot training to the stage of solo and somewhat beyond; several officers and enlisted men took various of the courses and flew series of missions at bombardier, navigator and radar observer schools; and a number of officers and enlisted men went through the complete course of flexible gunnery training and received ratings as aerial gunners. At a somewhat later date there were added to the staffs of the projects concerned with bombardier and navigator research psychologists who had first worked in those areas as civilians and who had then entered the Army Air Forces and gone through the complete course of aircrew training for the specialty under study.

An interesting project in both job and man analysis was undertaken by the two officers who were mentioned in the preceding paragraph as having taken a part of the course of pilot training. These men shared not only the training but also the life of the cadets. They lived in cadet barracks, ate at cadet mess, participated in cadet "bull sessions," and became insofar as possible a part of the cadet life at the post. Various indications were available to show that they had been fully accepted by the cadets and admitted to their confidence and fellowship. Under these circumstances it was possible to make intensive observations, not only of the process of learning to fly, but also of men learning to fly—their strain and tension, their focal points of difficulty and the relationship of their flying problems to the men and their background.

Studies of the men were based largely upon participant observer procedures in which the officers entered into conversation with the men singly and in groups, and subsequently recorded the gist of the discussion. These informal procedures were supplemented at a later stage by more intensive and systematic interview procedures. There were available as data not only the observations and interviews obtained in the training situation, but also all the test scores and other background information about the men which had been obtained at the time of their classification testing. The whole program was planned to provide a maximum of insight into flying training as it was experienced by the cadet in the army milieu.

The experiences of the personnel who received training as indicated in the previous paragraphs were undoubtedly of great value to the psychological research program. The value is most clearly manifest in connection with the construction of tests of achievement and proficiency, since it becomes extremely difficult for a psychologist without extensive training in the subject matter in question to do acceptable work of this type. The experience was also of value in connection with job analysis as a source of insights into the functions to be measured in selecting personnel for these particular jobs. Though some difficulties are still involved in making generally available to others who have not had the experience the insights of those who have received the training, this is much reduced when the job experience is had by psychological personnel and the report is prepared by individuals with that type of training.

The question of how far psychologists should go in mastering the particular specialty for which they wish to develop measures of proficiency or aptitude is one which raises a broader issue. Fundamentally it is a question of whether personnel psychologists, in addition to a background of experience and highly technical training in appropriate psychological techniques, should also be expected to master the specialties to which they apply these techniques. In general, it may be questioned whether such a philosophy of dual or multiple specialized training will be an efficient utilization of time and effort. The alternative in the Aviation Psychology Program was to draw heavily upon the background and experience of specialists in the various aircrew jobs. From a practical administrative point of view it was found that the utilization of and inclusion of operating specialists in the development of measuring instruments could result in a more wholehearted acceptance of these instruments by the operating organization. It may well be, therefore, that, in the long run, more practical contributions to the problems of the operating organization will result if the psychologist works essentially as a psychologist in close cooperation with specialized personnel, supplementing

systematic observation on his own part by the detailed knowledge of specialists in the field, than if the psychologist attempts first, at the expense of considerable time and effort, to master the specialty and then to proceed in his developmental work in considerable independence of the personnel in the organization which purports to benefit from his services.

TEST VALIDITIES AS A SOURCE OF JOB ANALYSIS

Although validity data become available too late in the cycle of test development to be of initial use in providing an understanding of job requirements, they may ultimately be a valuable source of insight into the abilities required for the job. Validity data provide an objective check upon initial hypotheses as to job requirements. As a considerable array of validity coefficients for different types of tests becomes available for study, together with the test intercorrelations, a good deal of insight into the factors related to that criterion may be obtained from an examination of the correlational data. This insight may be further refined by factor analysis procedures, in which both test and criterion correlations are included in the analysis. Studies of the classification test battery and the criterion of pass-fail in primary pilot training indicated, for example, that the factors identified as "mechanical," "space relations" and "aviation interest" had the highest validity while "verbal," "numerical" and "reasoning" had substantially zero validity. For navigation training, high validities were found for "numerical," "space relations," "science education" and "reasoning" while "coordination," "aviation interest" and "visualization" had near zero validity.

Analysis of validity data is valuable in clarifying aspects of the criterion which are already measured by existing tests. This may contribute to the improvement of a test battery by indicating factors for which improved and purified tests are needed. Tests may be found which were valid in combinations or for reasons not suspected at the time the job was originally analyzed and the tests constructed, and thus the concept of the job may be extended. However, analysis of validity data is limited to the factors which are in some measure included in those tests which were developed on the basis of the original job analysis. Within the scope of the original battery of tests, analysis of test validities and intercorrelations serves to check and refine the original job analysis, but these statistics do not provide a basis for extending the job analysis to new and virgin fields.

USE OF JOB ANALYSIS RESULTS

Previous sections of this chapter have discussed the sources of job information. It is now appropriate to give some consideration

to the procedures for making use of this information. First of all, of course, the activities involved in studying the job by one of these methods tended to provide the particular psychologist making the study various hypothesis, precisely or vaguely formulated, as to functions important for success in the particular aircrew duty studied and perhaps even ideas as to test procedures by which these functions might be measured. These were the informal, unanalyzed results of such a study.

In the second place, as indicated at the beginning of the chapter, somewhat formalized job analysis reports were prepared in the series of Analysis of Duties Bulletins, and in certain of the Research Bulletins which served as the mechanism for reporting various types of research studies. These reports supplied the reader, in most cases, a fairly detailed statement of what the man was required to do in the job in question. They then typically went on to propose a list of functions which appeared to the analyst to be important for that aircraft assignment. There was no uniform system of format or of categories in presenting job descriptions, and no uniform procedure was developed for extracting and presenting job analysis results. No uniquely satisfactory set of categories is known to exist in terms of which the analysis of any job may best be cast. The analyses were formulated as best the writer could, and made available to stimulate hypotheses for testing in the mind of the reader.

EVALUATION OF JOB ANALYSIS PROGRAM

Job analysis in the AAF Aviation Psychology Program was characterized less by novel contributions to technique than by the extensive exploration of familiar job analysis procedures. The job analysis was carried out under considerable pressure of necessity to get a test battery into operation in the shortest possible time. As a result, in the initial stages job analysis leaned heavily upon available records and reports and upon second-hand experience of other individuals. This was progressively supplemented by more and more direct personal participation by psychological personnel in training activities, with a corresponding increase in thoroughness of knowledge of the various aircrew specialties. In the sum total, the time devoted by psychological personnel to direct and vicarious experience of the job activities was very considerable. However, test construction was at all times in the program spread over a large number of participating workers. Most of these had to rely throughout the program upon somewhat indirect and second-hand sources for the suggestions and insights in terms of which tests were developed.

It must be admitted that job analysis procedures were not satisfactorily systematized or formalized. They proceeded largely on

a common sense basis, utilizing what information was available. Interpretations based upon information so acquired were on a highly subjective level depending entirely upon the insights of the individual research worker. No satisfactory formal framework was worked out as a guide to making job analyses or drawing interpretations from them. It may be that no such framework of procedures or categories is desirable or even possible; in any event, it was not achieved.

CHAPTER THREE

The Invention and Refinement of Aptitude Test Forms

Once a job analysis had been carried out and certain clues had been obtained as to the functions necessary for successful performance in a job, the Aviation Psychology Program then moved on to the task of inventing and developing tests to measure those functions. The task at this point was to translate each concept of a psychological function, as it had been abstracted from a particular job situation into a practical series of testing operations and then to refine those operations in the light of preliminary experience and trial so that they corresponded to the function as adequately and as accurately as possible. We shall consider first the steps taken in the aircrew testing program to foster the invention and formulation of experimental tests and then the procedures for test development and refinement.

FORMULATION OF RESEARCH TESTS

The aircrew selection program undertook from the first to foster the suggestion and discussion of all types of test ideas from whatever source. Since not only the officers in the program but also a large number of the enlisted men had had substantial amounts of psychological training, procedures were set up to encourage suggestions of ideas for tests by both officers and enlisted men. A Test Idea Form was prepared, which was revised from time to time, indicating the types of factors which should be covered in reporting an idea for a test, and personnel of the psychological program were encouraged to submit test ideas which they deemed worthy of development. Originally, and during much of the aircrew program, test development research was concentrated in three Psychological Research Units at classification centers, a Perceptual Research Unit at Headquarters AAF Training Command and the Department of Psychology, AAF School of Aviation Medicine. In each case test ideas submitted by some individual in one of these units were reviewed by the officer in charge of the unit and submitted to a central headquarters for review, assignment of a code number in a single over-all coding

system, distribution to other units of the program, and criticism and evaluation in terms of priority for development.

The Test Idea Form was planned with a view to stimulating suggestions for tests and at the same time stimulating critical thinking about those suggestions by the individuals initially responsible for making them. The form called not only for a description of the test procedure but also for a rationale for the test. The rationale was to indicate the functions which it was believed that the test should measure, the reasons for believing that the test would measure these functions, and the basis for thinking that the test covered functions not already adequately covered in previous test development.

The flow of ideas for tests stimulated by this approach was considerable. In all, probably five hundred test ideas reached Headquarters and had code numbers assigned to them during the course of the war. Quite a number of these, of course, never proceeded beyond the idea stage. However, the number of test forms actually developed and tried out also reached into the hundreds. Many of these were developed in the course of systematic coverage of some area by test development personnel, but a number of interesting ideas were also received from individuals in all different assignments in research and testing units. The systematic solicitation of test ideas appeared to have values both from the point of view of individual morale and from the point of view of its fruits.

APPROACHES TO TEST DEVELOPMENT

In undertaking a program of test development, at least three types of approach may be recognized. The first will be spoken of as the "hunch" approach. By this we have in mind the case in which an individual has been meditating about flying training and, as he goes on with his work, has a notion for a test. The notion arises more or less in isolation, perhaps stemming from the originator's dissatisfaction with some part of the existing battery, from his interest in some particular trait which he believes to be important for aircrew, or from some particular experiment or apparatus upon which he had been working prior to his military career. It is not part of any over-all plan or systematic program of test development.

The second approach may perhaps be designated the job analysis approach. In this approach the individual proceeds from a systematic study of the job and listing of the functions called for by the job, and endeavors to make tests of those functions. In this approach, the functions are seen to a greater extent in terms of the job than in terms of a fundamental pattern of human abilities. In the extreme, this approach leads to a job sample type of test

in which the tester endeavors to reproduce in a miniature situation all the complex conditions of the job itself.

In the third approach to the development of an aircrew testing battery, the initial effort is to define a set of fundamental, independent categories or traits of human behavior and then to develop tests of these basic traits. From that point the problem becomes one of determining which of the traits are in fact important for aircrew success and building a battery in terms of tests of the traits found to be important.

The job analysis and trait analysis approaches represent opposite extremes of a continuum, rather than unrelated approaches. In practice, most test development research falls along some intermediate range of that continuum, emphasizing to some degree both the reproduction of the conditions of the job and the analysis of basic traits of human behavior. By the same token, different research workers in the Aviation Psychology Program attached different degrees of importance to the two ends of this scale. In theoretical discussions, at least, certain individuals emphasized the necessity of obtaining pure measures of simple elements of human behavior, almost to the exclusion of any concern about the particular complex of job functions. Other individuals insisted that the essence of such a job as flying a plane was the complex of simultaneous activities and adjustments, and that tests which concerned themselves with isolated facets of human behavior would inevitably prove unfruitful.

Through the convictions of research personnel spread out on the continuum between job analysis and trait analysis, test construction operated in terms of the "hunch" approach as well. Insofar as test invention was indulged in by a large number of individuals at a number of different stations, it was natural that many suggestions for tests represented isolated thoughts of particular individuals rather than elements in a comprehensive and integrated program. Many of these tests were developed, insofar as they seemed to possess merit, without being completely integrated into a rational plan for an over-all testing program. Appreciating the limitations of human insight in planning an inclusive, comprehensive program, it seems probable that the encouragement of isolated ideas, without regard to a total framework of test construction, was a sound procedure for extending the scope of test development.

Consideration of the factors entering into the decision as to whether to concentrate test development on tests closely related to the particular job or on tests to measure "pure" traits of human behavior involves one in the complete field of correlational analysis in relation to test results. These problems are discussed in some detail in Chapter 8.

MEDIA OF TESTING

There were three major test media developed for use in the aircrew classification battery. These were respectively printed tests, motion picture tests, and apparatus tests. The types of test materials are to be considered supplementary in that each possesses certain unique advantages as well as certain limitations.

The greater part of the testing time of each subject tested was at all times devoted to printed tests. These possess as their outstanding advantage efficiency in administration and practicality. They also, of course, provide for a maximum of uniformity in testing conditions from individual to individual. However, there seems to be a certain range of human aptitudes which cannot readily be adapted to testing with printed tests. The range of traits which can be measured by printed test techniques is not clearly defined, and with sufficient ingenuity it may be possible to develop effective printed test techniques for types of performance which had previously been considered susceptible only to individual testing. In the aircrew classification tests, an example of this was found in measures of the "spatial relations" factor. Analyses of some of the classification test batteries together with groups of research tests indicated that a large fraction of the pilot validity of several of the apparatus tests could be attributed to a spatial factor which they shared with several perceptual group tests. Much, though not all, of the valid variance covered by these apparatus tests could have been covered by the group tests. It is possible that group test procedures could be developed to cover the other valid variance of the apparatus tests with the resulting simplification of testing procedures.

It was the general guiding point of view of those supervising test development in the Aviation Psychology Program that printed tests should be used for all functions for which they could be shown to be adequate and that a fair amount of the effort spent on test development activities should be devoted to endeavoring to develop group testing procedures for measuring the traits of importance which were not readily measured by those techniques.

In spite of efforts to broaden the field of group printed tests there appeared to remain certain functions for which printed tests were not adequate. These included areas in which motion of the stimulus is a necessary feature, and in this case motion picture tests appeared to be uniquely well adapted to the testing problem. Printed tests are also poorly adapted to almost all areas in which the speed or coordination of motor response is a significant feature. In cases of this sort individual testing with apparatus seemed to be almost a necessity. Again, with printed tests it is almost impossible to devise procedures for timing accurately the

exposures to each successive stimulus or the rate at which the stimuli are presented to the subject. Either motion picture or individual apparatus techniques can be devised to deal with this type of situation. Finally, where accurate timing of single responses by the subject is required, individual apparatus testing would seem to be almost a necessity. A certain amount both of research time and of testing time, therefore, was devoted to apparatus and motion picture tests. Motion picture tests of aptitude were developed in the areas of judgment of speed of movement and direction of movement, paced perception and memory of spatially complex patterns, and synthesis of patterns and movements exposed piecemeal. Motion pictures were also used in the development of certain proficiency tests of navigation and bombing, being used in this case in order to duplicate more adequately in a test situation the complexities and sequence of the actual task required on the job. Apparatus tests were developed for various types of measures of speed of reaction, accuracy of movement, and coordination in complex motor tasks.

The limitations of motion picture and apparatus tests are in considerable measure practical ones connected with test construction and use. A first obstacle in the case of motion picture tests was the very considerable amount of technical skill required to produce an effective film. Though some photography, especially for preliminary forms, was carried out by aviation psychologists, for most of the production of motion picture tests it was necessary to rely upon the technical skill of professional studios. This called for an intimate cooperation between the test constructor, who knew what effect he was trying to achieve in his test, and the technician, who knew how to achieve it. Other problems arose in the actual conduct of testing, but these appeared not to be as serious as had been anticipated. Studies of seating position failed to show evidence that this factor was a significant determiner of score in tests excepting those which taxed the limits of visual acuity. The lighting requirements of enough light to mark answers by and yet little enough so that the screen image was sharply defined appeared to be met by quite a range of illuminations. However, seating and lighting are two problems which require investigation and some degree of special arrangements in almost any case when motion picture tests are used.

In the case of apparatus tests, problems centered around the very considerable investment in personnel and equipment necessary to carry out testing on an individual basis in a large scale program and upon difficulties in maintaining constancy of testing conditions between different copies of an apparatus and within the same apparatus over a period of days or even months. In this program the standard procedure was to prepare all apparatus

tests in units of four copies combined with a single control table. It was found possible for a single test administrator to supervise and record the results from the testing of four subjects at the same time. In this way the efficiency of the testing program from the point of view of utilization of manpower was very considerably increased and it was possible to process men through a single "line" of psychomotor apparatus tests at a rate of 100 or more per four-apparatus lines per day. The problems of apparatus variation and difference between copies of an apparatus will be discussed in more detail in a later chapter. That the problem is a genuine one appears abundantly clear. It also seems that it is not insuperable when adequate maintenance and control procedures are used.

DEVELOPMENT OF TEST FORMS

The general sequence for development of a new test involved somewhat the following steps:

1. Conception of the test idea.
2. Construction of experimental test items or an experimental copy of the apparatus.
3. Tryout of the experimental materials upon small groups.
4. Analysis of tryout data.
5. Preparation of a revised test form or apparatus.
6. In many instances, further cycles of tryout and revision.
7. Finally, administration to a substantial cadet population and, ultimately, determination of test validity.

It will be worthwhile to state in somewhat more detail what was involved in each of these stages of development. The statements which follow can be thought of only as a typical pattern and not as applicable to each individual case, because the sequence of test development varied somewhat from one test to another.

Conception of Test Idea

There is very little that can be said about the initial insight which suggests a test idea. About all that can be done is to refer back to the procedures for fostering test ideas which have been described in the previous chapter and in the earlier section of this chapter. On the basis of provisions for study of and circulation of reports of job duties and of the characteristics required for aircrew jobs, ideas for tests came into being, were discussed and reviewed by other persons available at the same unit and, where considered promising, were carried to the stage of item development or apparatus construction.

Construction of Experimental Tests

With the emergence and acceptance of the original test idea, work began on the production of a workable test form. In the case

of printed tests, this involved the devising of instructions and test items and the construction of a sufficient number of the latter. In the case of motion picture tests the general script for photography had to be produced. In the case of apparatus tests a pilot model, often in rather rough and tentative form, was constructed for tryout. Typically, at this stage in test development the test items or the test apparatus were tried out on a number of other individuals in the testing unit in order to determine the adequacy of instructions, the general difficulty level of the testing procedure as it had been set up, and the distribution of responses that was obtained. Revision of the more outstanding weaknesses became possible in terms of the insights gained from testing unit personnel and, in part, by the suggestions of those somewhat sophisticated individuals as they were tested.

Experimental Tryout

When the test had been polished somewhat and a complete test form or working apparatus had been prepared, it was then usually administered to a small group of subjects under the standard testing conditions of classification testing. That is, the test was introduced into the classification battery at the end of the testing sequence and the subjects took it much as if it were a classification test. Typically, a sample of two or three hundred cases was tested at this stage.

Analysis of Tryout Data

Data from the preliminary tryout described in the preceding paragraph were then subjected to various statistical analyses to determine the distribution of scores on the test, the reliability of the test and, in the case of printed and motion picture tests, the characteristics of the individual test items. A distribution of test scores was typically obtained and a mean and standard deviation were computed. Several types of indices of reliability were computed in different cases. For some types of tests correlations between score on odd and even-numbered items were used. In many cases, however, the tests were speeded and it was therefore not meaningful to compute an odd-even reliability coefficient. In these cases the test blanks were usually so constructed as to consist of two comparable sections, each one separately timed. This became essentially an administration of two equivalent forms, one immediately after the other. Reliability was then determined from the correlation between the two halves of the test. Many apparatus tests were administered with separately timed trials and in these instances reliability was computed by correlating trials or groups of trials with each other. At this stage analysis of the single items of a test was concerned with the difficulty of the component items

and their internal consistency. Items were selected with a view to getting those which provided differentiation among cadets tested and, where the test was of a single, presumably homogeneous function, items which were consistent with performance on the rest of the test.

Preparation of Revised Test

In the light of information obtained from analysis of preliminary testing, the test form was then revised. Unsatisfactory items were eliminated or rewritten, time limits and scoring formulae were adjusted, testing conditions changed, and so forth, as seemed indicated on the basis of the preliminary results.

Further Cycles of Revision

In many instances the revised test form was again administered to a group of subjects for further statistical analysis in order to determine whether the deficiencies observed in the earlier form had been removed and whether the new, revised form possessed satisfactory characteristics of difficulty, reliability, etc. In some instances, several cycles of this sort were required before a test reached a form which seemed satisfactory. This cycle of try-out and revision was somewhat time consuming, and some question may be raised as to how much of it represented an economically sound investment of time, in view of the urgent need for getting results under the pressure of the wartime military situation. Refinement of this pre-validation level is limited to considerations of difficulty, internal consistency, and intercorrelations. Since quick validation is an urgent need in a wartime situation, less of the preliminary exploration is justified then than in a continuing long-time research program.

Validation Testing

When the test finally reached a satisfactory stage of development, it was then administered to larger groups of subjects in order that data might be provided in terms of which to validate it. Validation testing was carried on until a large group had been tested. Test data were then put aside until most of the individuals had completed training, and were then subjected to validity analyses as described in Chapter 5. Certain practical problems in validation testing are discussed in a following section.

TECHNICAL PROBLEMS IN CONNECTION WITH ITEM ANALYSIS

Item analysis of an experimental form of a test served the purposes of determining each item's difficulty and its correlation with total score on a group of items. In addition, item counts for the

separate response alternatives for an item provided diagnostic information on the effectiveness of the different misleads and permitted editorial review of those which attracted no responses or failed to discriminate.

Item difficulty was defined as the percent of the group knowing the particular item. This was in some cases determined from the total group, but more often it was estimated using only the upper and lower 27 percent of the group. The policy with regard to omissions and correction for guessing was not uniform, but the most generally accepted procedure was to eliminate from consideration those individuals who did not attempt the item or some subsequent item in the test and to correct for guessing by the formula

$$\text{Difficulty} = \frac{\text{Rights} - \frac{\text{Wrongs}}{n-1}}{\text{Number attempting}}$$

where n is the number of choices for each item.

Determination of the relationship of item to total test score presents a number of problems. On the one hand, there is a problem in determining when it is appropriate to carry out an internal consistency analysis. A second problem that often arises is that of deciding, in the case of a complex test, what constitutes the most meaningful total test score against which item analysis should be carried out. The third problem involves determination of the most appropriate statistic to be used as an index of item internal consistency. These problems are considered in the following paragraphs.

Certain test blanks, such as biographical data blanks, interest inventories, general information tests, and the like, are not planned as homogeneous measures of any single function. The blank merely serves as a convenient document to bring together a number of possibly useful items which are related only by a certain general similarity of form or testing procedure. In the case of an instrument such as this, analysis in terms of internal consistency seems to be beside the point. The essential analysis is validation of the separate test items. When item validities have been determined, it may then be appropriate, of course, to determine the correlation of each item with the other items (and with other tests used for classification), but the purpose here is to minimize rather than to maximize correlations. That is, each item becomes essentially a separate test, valid in its own right, and is useful in proportion as it is independent of the other valid tests. Internal consistency analyses were not ordinarily carried out for materials of this type.

In the case of other types of tests, a problem arises as to what should be taken as the total score against which item analyses are

to be carried out. This problem is especially likely to arise in achievement tests. The typical achievement test covers the rather varied content of a course of training (such as bombardier, navigator or radar observer training). It is often made up of a number of subtests or sections, each covering a more homogeneous segment of the total range of material studied. The problem becomes one of choosing between score on the sub-section and score on the complete test as a "total score" for use in item analyses. Most theoretical considerations would appear to favor the use of the single part score. There is some reason to conceive of the part as homogeneous, and to desire items within a part that are internally consistent, or at least to evaluate the quality of construction of items within a part by their internal consistency. However, the correlation between the parts of a test, where the content of the parts is based upon an analysis of the training and of the duties to which the individual is subsequently to be assigned, is something to be investigated after the fact rather than something to be specified in advance.

The approved computational routine for determining item correlation with the total test score used the chart and table developed by Flanagan.[1] This procedure makes use of the percent succeeding on the item in the top 27% of the group for total score and the percent succeeding in the bottom 27%. A chart and table have been developed for estimating correlation in the total group from these figures. Use of approximately the top and bottom 27% has been shown by Kelley to yield the greatest precision in estimating the correlation when the total score is forced into a dichotomy rather than being treated as a continuous variable. The values resulting from this procedure are estimates of the correlation between the two underlying continuous variables, and are strictly analogous to a tetrachoric correlation coefficient.

In a good deal of computational work in item analysis, phi coefficients were used.[2] This statistic is a coefficient computed from a two-by-two point distribution by the same operations as a product-moment correlation coefficient. When the phi coefficient is based upon the total population, it is possible to relate it analytically to other correlational procedures. In a good deal of work, however, phi coefficients were computed from the top and bottom 27% of the group on total score. The relative standing of items is somewhat more reliably determined by using these percents, as in the case of the correlation coefficient, but the absolute values of the phi coefficients obtained from such a fraction have no known relationship to other existing statistics. The phi coefficient has

[1] Flanagan, John C. 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. Educ. Psychol., V. 40, 1939, pp. 674-680.

[2] Guilford, J. P. Psychometric Methods. New York: McGraw-Hill. 1936

the quality, somewhat questionable from the analytical point of view, that the values of the resulting coefficients are a function of item difficulty. The nearer the difficulty to 50%, the higher the values for the phi coefficient tend to be. The net result is thus for coefficients to tend to favor items close to the 50% difficulty level and to penalize those of extremely high or low difficulty. Though such a bias may not be undesirable from the practical point of view, from the point of view of analytical clarity, it would seem preferable to keep internal consistency and difficulty clearly distinct.

PRACTICAL PROBLEMS IN VALIDATION TESTING

Practical problems in validation testing centered around the questions of who and where to test and how many to test. "Who" and "where" were involved because of the desire to get appropriate groups at a time when their experience and motivation would be comparable to that of new applicants for aircrew assignment and at the same time to test them when there would be a minimum of attrition and delay in obtaining validation data. "How many" was involved because it was necessary to apportion a finite amount of testing personnel and testing time, and a finite number of subjects among the demands of an almost unlimited number of competing research tests.

Administration of research tests for purposes of validation typically took place at two different stages in the sequence of aircrew classification and training. Part of the administration of research tests was carried out at the time of classification testing with the battery of aircrew classification tests. Other research testing was carried out during preflight training, subsequent to classification testing and prior to entering upon flying training. Each of these times presented certain advantages and each had certain limitations.

The initial classification testing of aircrew candidates took place prior to their acceptance for and assignment to any type of aircrew training. The advantage of administering research tests at this time was that testing conditions were most nearly identical with those under which the test would be used if it were ever introduced into the classification battery. The chief drawbacks were delays before validation and attrition of the original experimental population. In the early days of the program cadets went fairly directly from classification into preflight school and then into flying training in the case of pilots, or advanced training in the case of bombardiers and navigators, so that delays in test validation were kept to a minimum. Later on, the AAF college training program was initiated and classification testing was accomplished prior to the college training period. This introduced a very

considerable lag between classification testing and assignment for training. Still later the college training program was eliminated but various types of on-the-line training were introduced in its stead, and the lag between classification testing and entrance into aircrew training continued to be rather great.

Initially, there was relatively little elimination from training on the basis of classification tests and almost all men, excepting those who were physically disqualified, entered into one or another type of aircrew training. At this time the program for pilot training was very much larger than the programs for bombardier and navigator training and, in addition, a large proportion (possibly half) of the quota for bombardier and navigator training was filled by eliminees from pilot training. As a result of these factors, validation tests administered at the time of classification tended to give a large yield of pilot trainees but only a meager yield of bombardier and navigator trainees. Later the disqualification rate because of low psychological aptitude was progressively raised. At the same time, or a little later, the proportion of trainees assigned to bombardier and, particularly, navigator training increased. The net result of this was a general loss of cases between classification testing and training, and a sharp reduction in the yield in the category of pilot trainees. This loss was synchronous with the introduction of delays between classification and training. The delays and attrition made classification testing an unsatisfactory locale for administration of research tests.

Because of the difficulty which had existed at all times in getting a sufficient group of navigator and bombardier trainees, because of the general attrition between classification and training and the extensive delays between these two stages in the last two years of the war, it seemed desirable to carry out at least part of the research testing with groups who had already completed their classification. This testing was done in most cases in preflight schools with groups undergoing preflight training and about to enter into the aerial phases of their training. The emphasis in research testing shifted progressively from the classification situation to the preflight school situation.

Carrying out research testing at the preflight school level rather than as a part of classification testing had certain obvious advantages and certain possible limitations. The advantages were that there was a minimum of loss between testing and validation either in time or in subjects. The chief disadvantage was the possible difference in testing conditions between the individual who was being tested for classification, who knew that his opportunity for aircrew training and his chance of getting the type of training he desired depended upon his performance on the tests, and the individual in preflight school, who had already completed his

classification testing and who understood that the tests which he was taking would have no direct or crucial effect upon his own future. It is quite possible that lack of motivation affected test performance at the preflight level but no adequate studies are available to indicate whether such an effect did in fact appear or whether it influenced validity coefficients subsequently obtained.

It must be admitted that the lag between test administration and validation results was a serious handicap to the progressive improvement of the classification test battery. The serious practical objection to validation administration subsequent to training assignment was that of military expediency. In a rapidly expanding training program, which attempted to turn out graduates as rapidly as possible, it was generally not feasible to obtain the necessary testing time during preflight or other stages of training.

The problem of the size of group to use for validation testing was a constantly recurring one in the Aviation Psychology Program. It did not seem possible to give any definite analytical answer to this problem. Clearly, other things being equal, the larger the group the better. The size group was limited by practical considerations of the size of population available for testing and of the demands of other research tests for time in the testing program. It was the tacitly accepted goal of the psychological program to have available for validation analysis a minimum of 1,000 cases for any research test for a particular aircrew category. Because of the several types of aircrew specialty for which classification was made and because of attrition from various causes between classification center and completion of training, it was necessary to test a minimum of some 2,000 cases in a classification center in order to get 1,000 cases who could subsequently be followed into pilot training. The proportion in any group assigned to bombardier and navigator training was generally so small that it did not seem feasible to get an adequate population of bombardiers or navigators from research testing in a classification center. Validation of research tests for bombardiers and navigators depended upon special testing, specially planned at the preflight school level, and excepting for such testing, it was generally true that the only tests for which validity data became available in adequate amounts for bombardiers and navigators were those tests which had been introduced into the classification test battery either in the initial design of that battery or because of their subsequently demonstrated validity for pilots.

Validation of tests for aircrew categories other than pilot, navigator, and bombardier was undertaken only at a relatively late period. Research tests for such specialties as radar observer or flight engineer were administered to groups who were about to enter training for that particular specialty. These were groups

who had already completed training and in some cases field experience in some other aircrew specialty. Interpretation of test validities based upon groups of this type presented particularly knotty problems, due to the several types of restriction which the groups had suffered from previous testing or training. These problems are discussed further in Chapter 5.

CHAPTER FOUR

Problems in Determining An Adequate Criterion

THE CRUCIAL ROLE OF THE CRITERION

Certainly the most fundamental and probably also the most difficult problem in the Aviation Psychology Program was that of obtaining satisfactory criterion measures against which to validate tests and evaluate variations of training methods. The criterion is absolutely central to any research program in testing or in training and in the Aviation Psychology Program investigation and development of criterion measures called for much research effort and insight. Other research can hardly proceed until a criterion is provided, and can be only as good as that criterion. There may be certain traits for which it is difficult to develop tests, or specific training problems in which it is difficult to formulate the desired training procedures, but these difficulties are limited to specific areas. Until some solution of the criterion problem has been reached, the progress which can be made in test development or training research of any kind for that particular aircrew specialty is very limited. Of course, research does not and cannot wait until a wholly satisfactory solution of the criterion problem is reached; in that case one would probably never do any research. However, some compromise solution of the problem of providing a criterion of success in performance of the task in question must be arrived at as a basis for any effective research program.

It is perhaps worth re-emphasizing that the criterion is as important to training research as it is in aptitude testing. To be sure, we cannot validate a particular aptitude test until we have a criterion of success against which to correlate test scores, and any further analyses which endeavor to combine tests into a battery will be dependent upon this validation. However, it is equally true that we cannot compare alternative training procedures until we have established some measures of the outcomes of that training which we are willing to accept as providing an index of success. As these two different areas of research in aviation psychology are explored, they each demand adequate criterion measures for the solution of their basic problems.

GENERAL PROBLEMS IN CONNECTION WITH CRITERIA

We have indicated that some solution of the criterion problem must be arrived at before further research can be effective. However, there are all degrees of adequacy of solution of this problem. In any given practical instance a number of possible criterion measures will usually suggest themselves, each of which has some degree of adequacy, less than complete, and some degree of practicality. Developing a practical research program will require the evaluation of and selection from among these possible criterion measures.

For purposes of discussion it may be profitable to differentiate three categories of criteria: ultimate, intermediate, and immediate. By the ultimate criterion is meant the final goal of a particular type of selection or training. For example, it might be agreed that the final goal of training Army Air Forces bombardiers was that they should under conditions of combat flying drop their bombs in every case with maximum precision upon the designated target. The ultimate goal for a career gunner might have been that he score the maximum possible number of hits upon attacking fighter planes. Such a goal is likely to be stated in very broad terms and in terms which are, in many cases, not susceptible to quantitative evaluation. Furthermore, it will usually not be entirely accurate to specify a single and unified ultimate goal. The bombardier had to fire a gun as well as drop bombs. The gunner had to cooperate effectively with other crewmen in the identification and selection of targets as well as firing his own guns. An absolutely complete ultimate criterion will probably be multiple and complex in most cases. Such a criterion is ultimate in the sense that we cannot go beyond it to look for any higher or further standard in terms of which to judge the outcomes of the selection or of the training.

In practice, ultimate criteria are rarely, if ever, available for use in psychological research. They may be completely inaccessible, but in the event that they are potentially available they are likely to be far away both in time and in space, confused by a number of other interacting factors, and difficult to express in usable quantitative form. In such a case we are almost inevitably thrown back upon substitute criteria which we judge, either in terms of our rational analysis or in terms of empirical evidence, to be related to the ultimate criterion towards which we aspire. These criterion measures we may designate as intermediate or, in certain cases, immediate criteria.

The term immediate criterion is used here merely to differentiate that criterion measure which becomes available most immediately and directly from other partial criteria which become avail-

able at various later stages in the course of training or of performance upon the job in question. For example, we may consider the immediate academic criterion in the case of pilot training to be graduation or elimination from preflight and the immediate flight criterion graduation or elimination from primary pilot training. These are the first major objectives of the training program and provide in at least a negative sense data on success in combat flying. Certainly, by definition, the man who is eliminated from training at the primary level cannot be a valuable and successful combat pilot; he will never be a combat pilot at all.

Examples of intermediate criteria in pilot training would be graduation or elimination from basic, advanced, or transitional training, score in fixed gunnery at the transitional or operational training level, or ratings by supervisory personnel either in advanced training or in the theater of combat operations. Even in the case last mentioned, the ratings remain intermediate rather than ultimate, because we are not ultimately concerned with how a man will be rated by his superior officers but rather with how well he will actually perform in the crucial situation for which he has been trained. All immediate and intermediate criteria remain partial, therefore, since at best they give only an indication of or approximation to the ultimate goal towards which our selection or training is directed. A research program must start at an early stage to analyze the available immediate and intermediate criteria in order to determine as far as possible the adequacy of each as an approximation to the ultimate criterion towards which the program of selection and training is aimed.

The ultimate criteria of success in any duty are always determined on rational grounds. There is no other basis upon which this choice can be made. In some cases, agreement in selecting the performance records which would serve to define the ultimate criterion may be arrived at quite readily; in other cases the process of defining ultimate criteria may involve prolonged and agonizing soul-searching. Once agreement has been reached on ultimate criteria, it may be possible to carry out part of the evaluation of intermediate criteria in terms of empirical data on their relationship to the ultimate criterion. This will ordinarily be carried out for certain special experimental groups. For these, criterion data at various stages will be collected and correlated with those measures which have been selected as being the best representation of the ultimate criterion. Those intermediate criterion measures which show high correlation with the ultimate criterion for this special group will then be selected for routine use in the many groups upon which test validation and training research are carried out. However, limitations of time and of the availability of data will often require that intermediate criteria be evaluated in

terms of their rational defensibility and in terms of their internal statistical characteristics, rather than in terms of their relationship to ultimate criteria.

In the case of the aircrew duties with which the Aviation Psychology Program was concerned, the ultimate criteria of performance lay in combat. It was very difficult to obtain combat criterion data which were satisfactory either in quantity or quality. The field of battle is not an easy one in which to glean psychological data. The control of experimental conditions in actual warfare is, to say the least, very far from ideal. Moreover, there was necessarily a long time interval between the date upon which a group of men were tested and the date that criterion data matured for them. It was only relatively late in the war that psychologists were able to get to the combat theaters and obtain performance records of men who had previously been tested. Even then, the records which were available were incomplete, lacking in uniformity, and organized in such a way as to be very difficult to use in psychological work. It was necessary, therefore, in most of the research work in classification test development and investigation of training procedures, to validate tests or procedures against immediate and intermediate criteria which had not themselves been correlated with the ultimate criterion.

Among the intermediate criteria of varying degrees of immediacy various internal analyses were possible and were carried out. Correlations were determined among criteria at successive stages of training, such as primary, basic, and advanced pilot training. Different types of criteria, such as grades and ratings, were correlated with one another. Data with regard to the reliability of single criteria were obtained where the criterion was of such a nature as to permit a reliability analysis, i. e., where two independent estimates of the criterion could be made. These analyses indicated the extent to which the intermediate criteria satisfied certain necessary conditions for validity in terms of more ultimate criteria, even though they left unanswered the question of the actual correlation of these measures with the ultimate criterion.

In the analysis of the relationships among intermediate criterion measures, it is not always clear which measure is being tested and which is serving as a standard. Neither measure is ultimate, and it may be that neither has a clearly better rational defensibility than the other. The investigator ordinarily thinks of one rather than the other measure as being the standard, on the basis of factors such as nearness in time to the ultimate performance, acceptance by operational personnel, directness of apparent relationship to the ultimate task, and the like. However, the discrimination may not be at all clear-cut in many cases and so the study

of interrelationships may be thought of as throwing light on each of the criterion measures. In general, high correlation between different intermediate criterion measures will tend to strengthen the rational basis for accepting either of them as a useful criterion, since each will then receive some support from the rational justification of the other. Lack of correlation may tend to weaken one or both of the measures, except insofar as they measure distinct aspects of performance for which there is no rational basis to expect intercorrelation.

In some cases, when standardized aptitude or achievement tests are correlated with intermediate criteria, there may be some question as to which is being evaluated, the criterion or the test. That is, specially constructed achievement tests and even some specially constructed aptitude tests may have a sufficiently good rationale as measures of the job under study so that they appear as acceptable as certain available training measures. This tends to be true for selection tests in proportion as they become job miniature tests. In the Aviation Psychology Program, for example, a gunnery sighting test was developed at the School of Aviation Medicine which seemed to some observers to represent as close an approach to the task of combat firing as almost any situation presented in training. When this situation arises, it becomes, of course, rather futile to endeavor to "validate" the test against existing training criterion measures.

EVALUATION OF CRITERION MEASURES

A criterion measure must be evaluated, in the last analysis, by whether it does in fact provide a score that correlates highly with the theoretically perfect ultimate criterion. The necessary and sufficient conditions for such correlation are that the criterion measure have relevance to the ultimate criterion, reliability, and freedom from bias. The necessity for and conditions of these requirements will now be discussed in more detail.

Relevance

The quality which has here been designated "relevance to the ultimate goal" is the first essential of a criterion measure. A criterion measure is relevant insofar as the task requires of the individual the same knowledges and skills and use of the same basic aptitudes which will be required for performance of the ultimate task. Theoretically it would be possible to determine the relevance of a criterion empirically by its correlation with the ultimate criterion. In practice, the *complete* ultimate criterion is never available, and near-ultimate criteria may be extraordinarily difficult to obtain and may be unsatisfactory in other regards. The result is, as indicated earlier, that the relevance of a particular criterion measure will usually have to be estimated on rational

grounds with only limited help from empirical data. The problem is analogous to that of determining what should go into an academic achievement test. Rational analysis must be relied upon to a very large extent in determining what the goals of instruction are and consequently what is appropriate content for the test. The adequacy of the rational evaluation of criteria will depend upon the intimacy of the analyst's knowledge of the ultimate goal on the one hand and of the immediate criterion measure on the other, and upon the basic sagacity of the analyst.

Relevance is the absolutely fundamental requirement in a criterion. Insofar as at all possible, it is important that *all* systematic variance in the criterion measure be relevant variance. If the criterion measure possesses any appreciable amount of irrelevant non-chance variance (or worse yet, variance negatively related to the ultimate goal), it is entirely possible that a systematic differentiation in research results, whether in selection or training, may be based entirely upon this irrelevant variance. That is, it might be possible to develop selection procedures which would give quite a good prediction of a moderately relevant intermediate criterion and yet have exactly zero relationship to the ultimate criterion. This would be the case if the prediction was based on that fraction of the systematic variance in the intermediate criterion which happened to be irrelevant to the ultimate goal. Thus, a vocabulary test might give a good prediction of grades in gunnery school, and yet be entirely useless in selecting good combat gunners. In this case, a selection procedure would have been developed which appeared superficially to be quite successful, and yet which had absolutely no fundamental value. This possibility is always present when the criterion being used is only partially relevant.

Reliability

A necessary but not a sufficient condition for correlation between a criterion measure and the theoretically perfect ultimate criterion is that the measure have *some* reliability. That is, the reliability must be greater than zero, because if the reliability of the measure is zero it cannot possibly correlate with *anything*. Evidence with regard to reliability is primarily statistical. In the statistical evaluation of a criterion measure the first essential is to get evidence which would require the rejection, at a satisfactory level of confidence, of the hypothesis that the reliability of the criterion is zero. High reliability in a criterion is not critically important, though it is convenient. Low reliability in a criterion will merely serve to attenuate all its relationships with other measures and also the effect of special experimental variables. It cannot produce systematic stable relationships as may happen for measures low in relevance. Insofar as low reliability is due to

random, chance factors, it can produce only a weakening of relationships. It is possible to compensate for this attenuation of data by increasing the size of the experimental population, and thus determining the values of all statistics with greater precision.

Low reliability is occasioned by inconsistent performance by the subjects being studied and by fluctuations in the external conditions and definition of the task. These may be called intrinsic and extrinsic unreliability respectively. Intrinsic unreliability can be reduced only by increasing the sample of behavior included in the evaluation, that is, by doubling the number of rounds fired, bombs dropped, check flights flown, etc. This has the effect of reducing the proportion of chance variance in the total.

Extrinsic unreliability will also be reduced by extending the sample of behavior observed. It may also be reduced by controlling the external conditions. These external conditions include both conditions influencing the performance and conditions influencing the observation of the performance. Conditions influencing the performance are such factors as weather, equipment, other personnel entering into the situation, and the exact definition of the task. Reducing variation due to these factors, insofar as it is feasible, presents a complex administrative problem of maintenance of equipment, scheduling, briefing of personnel, and the like. The degree of control which can be achieved is limited by the extent to which schedules, equipment, and other personnel involved can be kept at a uniform standard within the practical situation of a large-scale training or operating program.

Conditions influencing the observation of behavior are the preciseness of definition of the behavior to be observed, simplicity of the behavior, degree to which the behavior is overt, amount of aid provided by instruments, and opportunities which are provided to observe the behavior. Efforts to improve reliability of observation will be devoted to breaking the behavior down into simpler components which can more readily be observed, defining the behavior to be observed as precisely as possible for the observer, providing a maximum of physical opportunity for the observer to see the behavior in question, providing mechanical aids and records, and the like.

Freedom from Bias

Bias is a condition which may operate to reduce either relevance or reliability or both. Its effect depends upon whether the bias happens to cut in a random fashion across the groups being compared or whether it effects the different groups in a selective way. For example, consider the bias which is represented by differing standards for elimination in different primary pilot schools. On the one hand, we might be interested in studying the relationship

of elimination rate to score on a particular test. If test scores are randomly distributed among the schools, the effect of school differences in elimination standards becomes a random rather than a systematic one, and serves merely to attenuate slightly any true relationship between test score and the graduation-elimination criterion. In another case, we might be interested in comparing two different training procedures, one of which was in effect in one school and one in another. Here, clearly, any difference in standards for graduation becomes a systematically biasing influence and will produce consistent differences between the two groups being studied which cannot be differentiated from any effects of the different procedures for training.

Bias may arise whenever sub-groups of a total population are evaluated in systematically differing ways. The sub-groups may represent those taught by a particular instructor or group of instructors, those in a particular school, class, command or combat theater, those well known as opposed to those only slightly known to rating personnel, and the like. Bias may arise within subjective evaluation standards or within external conditions. Bias is a much more serious matter than low reliability because it is always possible that it may affect systematically the comparison in which we are interested and thus produce spurious results. Bias is not as universally undesirable as low relevance, since it may be possible so to design the experimental situation that the biasing factors are randomized with regard to the factor being studied. However, the existence of bias always renders experimental results somewhat less secure since the procedures for randomization may turn out to have been imperfect. Bias becomes a matter of acute concern especially in studies of training procedure. In those cases we almost universally have two or more discrete groups to be trained, and administrative convenience almost always requires that they be physically separated. This opens the way for biasing factors to enter in, and it is in these situations particularly that lack of susceptibility to bias becomes a valuable characteristic of a criterion measure.

TYPES OF CRITERION MEASURES

The criterion measures used in the Aviation Psychology Program were of two broad types, specific evaluations of a limited unit of performance and summary evaluations of a total phase or large unit of training or operations. Each of these has its advantages and limitations, and each has its place in a program of criterion development. Specific evaluations of a limited unit of performance have the great advantage that they make possible relatively exact statement and specification of the criterion situation and of the conditions of observation of the behavior in which the research

worker is interested. This same degree of specification and control of the nature and conditions of the evaluation can probably never be achieved in the summary evaluation of an extended period of training. On the other hand, the summary evaluation covers a scope, in terms of amount and variety of behavior, which cannot be compressed into a limited test or observation period.

Study of specific evaluations was in large measure concentrated on procedures which were developed by aviation psychologists for the particular purpose of providing measures of proficiency. In contrast, many of the summary evaluations studied were those which were already recorded as part of the administrative routine of the training program. In the following sections we will first consider the types of specific evaluation which are possible, and explore the limitations and advantages of each as these exhibited themselves in aircrew training. We shall then turn to the types of summary evaluation, examining how these are derived from specific evaluations, and how their advantages and limitations are related both to the specific elements from which they are compounded and to the manner of compounding those elements.

SPECIFIC EVALUATION OF A LIMITED BEHAVIOR UNIT

The specific evaluation of behavior within a limited behavior unit may be subdivided as it is concerned with the evaluation of knowledge and information about the duties or as it is concerned with performance of the duties of the job.

Evaluations of knowledge and information take the form of the traditional "test," and the usual assortment of tests of varying degrees of objectivity and technical excellence is encountered. Measures of performance will bear further subdivision according as the performance is judged by means of an objective record of the performance, by means of subjective scoring of items of the performance, or by subjective rating of the performance as a whole. These three categories are not entirely separate and discrete, but represent identifiable points on a continuous scale from objective to subjective. At one extreme, the behavior itself yields a persisting record, and the observer enters in only to transcribe or score the record. Since the record persists, any necessary amount of time can be devoted to scoring it or any necessary repetition or verification of the scoring can be carried out to make sure that the inaccuracy or bias of the observer is reduced to an insignicant amount. The middle point on the scale, subjective scoring of items of performance, is encountered whenever the behavior leaves no lasting record but when specific segments of behavior may be evaluated as they occur in terms of such relatively simple and analytical judgments as amount of shift in instrument reading, angle of plane in a turn, position of landing on

a field, and the like, or the occurrence of such behavior as carrying out the steps of a preflight check. At the other end of the continuum we encounter the relatively unanalytical rating of the complete sequence of behavior. This is well illustrated by the grade recorded on a pilot check flight. This grade represents the synthetic evaluation of the complete segment of behavior occurring in a ride of perhaps an hour. It is an unanalyzed clinical judgment, in which it is no longer possible to identify specific behavior units and for which it is impossible to determine the way in which the items of behavior were weighted in the final composite judgment. However, the behavior upon which the judgment was based *is* still restricted to that shown in a specific, delimited segment of time.

To recapitulate, we find a variety of performance measures. These differ in the degree to which the evaluation is mediated by the observer. At one extreme, the behavior leaves a permanent record which may be scored or evaluated at leisure, thus making the evaluation on the one hand easier and on the other hand repeatable, so that the evaluation can to a very large extent be freed of the influence of the observer. Toward the middle of the scale are encountered situations in which the behavior leaves no lasting record, but where the behavior is still analyzable into rather simple units and where the necessary observation can be defined in terms of observable readings of instruments, occurrence or non-occurrence of simple items of behavior, and the like, which require only relatively direct perception on the part of the observer. These are mediated by the observer in the sense that his on-the-spot evaluation of the behavior provides the only available record, but the judgment can be defined in such simple terms that we may anticipate that individual standards of judgment will be of minor importance. As the situation becomes more complex, and the required observation more difficult to define exactly, we may anticipate that the medium of the observer will become of more and more importance until it reaches a maximum in the undefined rating of the complete behavior sequence. The reader will recognize in this discussion an elaboration of the conditions making for *objective* evaluation at one extreme and *subjective* evaluation at the other. The important points to remember are that objectivity-subjectivity is a continuum, and that the conditions making for objectivity are persistence of the trace of the behavior and simplicity and precise definition of the phenomenon to be observed.

Evaluation of Knowledge and Information

A background of related knowledge and information was recognized as having significance for success in most of the aircrew

duties with which the Aviation Psychology Program was concerned. However, since all the aircrew assignments involved primarily "doing," tests of "knowing" were generally considered to be somewhat peripheral to the main current of achievement. Knowledge and information had one great advantage in that they could be evaluated on the ground, making use of printed tests. A number of such tests were developed by aviation psychologists for aspects of almost all of the aircrew duties with which the Aviation Psychology Program was concerned. Thus, printed proficiency tests were developed for pilot, navigator, bombardier, radar observer, flight engineer and gunner, all of which involved in some degree knowledge about the job in question. Many of these tests were also, in a degree, performance tests in that they required the subjects tested actually to carry out certain of the sequences of computation required for figuring position, altitude, fuel consumption, and the like. Since in this area it was possible to prepare, through joint efforts of aviation psychologists and technical specialists, proficiency tests which met professional standards of test construction, it rarely seemed desirable to make direct use of tests which had been prepared by training personnel. These were, of course, represented in any over-all summary evaluations of which they were a part. Proficiency tests involving knowledge and information were used to some extent as criteria against which to validate selection procedures. They found further use as selection procedures at advanced stages of training, to select personnel for special duty such as lead crew training. Since the development of printed proficiency tests by aviation psychologists did not get under way until relatively late in the war, use of these measures was limited to the last year or so of hostilities.

Standard printed tests of knowledge and information were quite satisfactory from the standpoint of reliability and freedom from bias. As has been indicated above, it was on the count of relevance to the ultimate criterion that their value seemed limited.

Evaluation of Performance

Techniques for evaluating performance of the major features of the job were sought as the basic type of criterion material. These performance measures were needed for three distinct types of functions: to serve as criteria for validation of aptitude measures, to serve as criteria in experiments on modifying training, and to serve as selection devices in picking the most proficient individuals for such special assignments as lead crew training. Although the categories are not entirely distinct, it will help to organize the discussion to present these materials under the three subdivisions of objective performance scores, subjectively scored

job samples, and rated job samples. These three categories have been discussed above.

Objective Performance Scores

The objective performance score is potentially the ideal criterion for much of selection and training research. Since the ultimate criterion is almost universally a performance, in the case of aircrew, performance under the stress of combat, records of appropriate types of performance under experimental conditions seem to have a good deal of direct relevance to that ultimate criterion and to be thoroughly reasonable criterion measures. For example, skill in tracking and framing an attacking fighter, as recorded by a gun camera, appeared to come as close to proficiency in the critical combat duty of a gunner as any measure that one could hope to get in an experimental situation. At the same time, the fact that the task leaves a permanent and objective record minimizes the possibility of observer unreliability or observer bias entering in to attenuate or prejudice conclusions. Insofar as the external conditions of the task can be completely specified and rigidly controlled, it is then possible to present a standard task to the subject at any time and at any place.

Because of the attractive possibilities which this type of measure presented, objective performance records were sought in every aircrew specialty with which the Aviation Psychology Program was concerned. On the one hand there were many situations in which special performance situations which yielded a direct record, were set up for research purposes and on the other the existing performance records of the training program were explored and exploited to the full. Experimental situations involved in some cases actual in-flight test situations; in other cases they involved ground tests or trainers. The flight situation appeared to involve most directly and completely the type of performance for which the individual was being trained. On the other hand, the actual conditions of flight added a number of additional complications to the problem of controlling the test situation. Three illustrations may be cited of specific flight tasks yielding objective performance scores which were set up for research purposes.

In a series of validation studies for flexible gunners which were carried out jointly by the Department of Psychology, School of Aviation Medicine, and the Research Division, Central School for Flexible Gunnery, the criterion of proficiency was accuracy of tracking and framing an attacking fighter with combat-type equipment during a series of aerial gun camera missions. The gun camera took motion pictures of the point at which the gun was aimed, and provided an objective record of the gunner's performance. Each gunner was tested with a series of attacks and

both bomber and fighter pilots had been briefed in order to have these attacks made in as standardized a fashion as possible from man to man.

As an investigation of the course of learning and the needed amount of practice in bombing, the Psychological Research Project (Bombardier) undertook, in mid-1945, a project involving bombing under experimentally controlled conditions. The record was the photographed bomb-strike on a standard desert target. Estimated values were used for certain cases for which no photographs were available. The experimental design involved control of airplane, pilot, bombsight, bombing altitude, length of bomb run, and a number of other variables. The plan was to carry a group of 100 students through a total of 450 bomb drops (three times the standard amount) and record the course of their improvement, but the end of the war led to the termination of the experiment after each man had dropped about 70 bombs.

A rather different type of objective record was obtained for navigators by using the logs of a series of formation navigation missions. These were initially used by the Psychological Research Project (Navigator) as the criterion for evaluating a dead-reckoning navigation trainer. The observations and calculations in each navigator's flight log provided an objective statement of where he believed himself to be at specified times during the trip. The use of the log as an objective performance measure required only that the conditions of flight be uniform for different navigators and that there be some way of knowing the *actual* flight course of each plane. The attack upon these requirements was to fly missions in formation, thus standardizing the flight for all men in the group, and to have in the lead plane two graduate navigators who provided standard values for the flight, against which the logs of the student navigators could be checked. A number of precautions were taken in order to obtain the maximum amount of standardization within each flight and from flight to flight.

On the ground, objective performance records were obtained on the one hand by some of the printed proficiency tests and on the other by various synthetic training devices. In a number of the proficiency tests part of the content required *doing* tasks which comprised parts of the job, as well as *knowing about* them. Of special interest in this connection were two motion picture tests which were under development at the end of the war. These were an effort to introduce into a classroom testing situation more of the elements of visual presentation and of pacing which characterized the actual flight situation. Aside, of course, from the fact that only certain of the more intellectual of the flight duties, such as the various types of computation, could readily be introduced

703815—47—4

into a printed test, the major criticism of printed performance tests was their artificiality.

Synthetic trainers provided a variety of objective performance scores. These trainers were encountered in greatest profusion in gunnery training, and it was in this type of training that they tended to yield performance scores, because by its very nature the aiming and firing of guns lent itself to the recording of "hits." Though they varied greatly in this respect, most synthetic trainers seemed less directly relevant to the ultimate combat criterion and on those grounds the rational basis for accepting them as criteria seemed somewhat less satisfactory than was the case with flight criteria.

In a previous paragraph it was indicated that complete specification and rigid control of the external conditions is necessary if objective performance records are to fulfill their promise as the ideal criterion of proficiency in aircrew duties. It is at this point that the major limitations of this type of criterion measure lie. It is fundamentally extraordinarily difficult to obtain the desired degree of specification and control. The conditions surrounding the performance of an aircrew member in the air are enormously complex. They involve, first of all, all the conditions of temperature, visibility and turbulence which constitute the weather. Here, only a limited control is possible by restricting the times of day during which flights will be flown, or by cancelling flights when weather conditions are too unfavorable. Under the practical pressure of a restricted time schedule, even this amount of control may not be possible.

A second major group of factors which must be controlled are those dealing with equipment. Calibration and maintenance of the bombsight, accuracy of alignment of driftmeter, uniformity of compasses and airspeed meters, and so on for the many guides upon which the bombardier, navigator, or pilot must rely for the information which he uses to carry out his task, all influence how well an individual will score. Where the typical personnel error is only one or two degrees, mils, or miles per hour, a small instrument or equipment error may become a major determiner of the final result. Just the chance variation in physical characteristics among practice bombs, for example, might constitute a substantial part of the error in good bombing. These factors could theoretically be reduced to minor importance by perfect maintenance of equipment. However, research had to be done not under theoretical but under actual conditions of maintenance. Under these conditions, lack of uniformity of equipment became a major problem in the use of this type of criterion measure.

A third type of factor which entered in to complicate the evaluation of objective performance scores was influence of personnel

other than the individual being evaluated. This can be seen most clearly in the case of radar bombing. If accuracy of actual or simulated (photographed) bombing is being used as a criterion, to what extent is the end result attributable to the bombardier? The radar observer? The navigator? The pilot? The effort to evaluate the individuals in a particular job assignment is vastly complicated by the fact that the available score is in varying degrees a function of men in other assignments in the plane. If it is possible to rotate crew assignments, the influence of personnel in different positions may be isolated and separately evaluated. In this case, variance from other crew members becomes attenuating rather than systematic variance. When the crew remains as a unit, however, or when systematic rotation is not possible, it becomes impossible to determine to which member of the team observed crew differences are attributable. Score may also be influenced by personnel other than those in the plane with the man being evaluated. In flexible gunnery, for example, the task which is presented the gunner is determined both by the pilot of his plane and by the pilot of the attacking fighter.

The above types of factors represent the ones which it is most difficult to standardize even in a defined, experimental test situation. In the ordinary conduct of training, and to an even greater extent in combat, various other factors in the situation are unstandardized and make interpretation of objective performance records more difficult. These can be subsumed under the general category of lack of uniformity of the task. Variations in route, target, opposition, and the like introduce a large amount of additional chance variance into whatever scores may be obtained under these conditions.

In terms of relevance to the ultimate combat criterion, objective performance scores would generally appear to rate high. They would also appear to be generally satisfactory in terms of freedom from bias, if adequate provisions can be taken to randomize the disturbing factors referred to in the previous paragraphs. This randomization becomes very important whenever discrete groups are being compared. The point at which the adequacy of objective performance records is least assured is reliability. In practice, the sources of unreliability are so manifold that it is necessary to make critical inquiry in each case to make sure that an appreciable fraction of the variance in the resulting score is associated with the individuals being evaluated. Returning to the three examples cited in earlier paragraphs, for the study of gun camera firing missions, we find the following reliability data reported for 10 missions:

	Rho
Odd *vs.* even missions, 16 Sperry gunners, tracking error..	.26
Framing error	.78

	Rho
16 Martin gunners, circular error	.74
Percent hits	.60
1st *vs.* 2nd 5 missions, 16 Sperry gunners, tracking error	.68
Framing error	.75
16 Martin gunners, circular error	.56
Percent hits	.42

Reliabilities for the bombing experiment were reported as follows:

	r	N
Odd *vs.* even missions, 45 bombs excluding 1st 6 dropped	.05	94
Odd vs. even missions, 1st bomb of each mission period including 45 bombs after 1st 6 dropped	–.08	94
First *vs.* 2d half of missions, all bombs, missions as above	.16	94
First 6 bombs *vs.* following 45 bombs	.18	94
Six bombs dropped from 4,000 feet *vs.* 45 bombs dropped from 7,000 feet	.18	89

In the case of navigation formation missions, the most critical over-all value determined by the student is his position at key points in the flight. Reliabilities for error in this performance were:

	Av Rho	N
Group I:		
Mission A *vs.* Mission B	–.06	139
Mission A *vs.* Mission C	–.16	146
Mission B *vs.* Mission C	.01	151
Group II:		
Mission A *vs.* Mission B	–.10	80
Mission B *vs.* Mission C	.13	80
Mission A *vs.* Mission D	.01	80
Mission A+D *vs.* Mission B+C	–.03	80

Though the reliabilities of the gun-camera scores appear fairly satisfactory, the bombing and navigation reliabilities are quite low. The navigational values, at least, are entirely consistent with the hypothesis of zero reliability in the population. This lack of reliability was in part a function of variation in conditions between missions (within-missions reliabilities were considerably higher); in part it was probably a function of inherent inconsistency of individual performance in these complex tasks. In any event, low reliability is often the limiting factor in the value of objective performance scores.

Subjectively Scored Job Samples

This category is used to cover those situations in which the performance itself leaves no lasting record, so that it must be

evaluated as it occurs, but in which the evaluation can take the form of direct observation and scoring of limited and rather well defined units of behavior. The behaviors range from those in which a recording instrument could readily be substituted for the observer, if an instrument happened to be conveniently available, to those which require a moderate amount of synthesis and interpretation by the observer. In the Aviation Psychology Program, this type of evaluation was represented by instruments variously designated as "phase checks," "performance checks," "objective scales of flying skill," and the like. They were prepared in great numbers, for there were many types of behavior which left no permanent performance record and these procedures seemed to provide the nearest approach to the objectivity of such a record.

A number of examples of this type of evaluation may be cited. At the Psychological Research Project (Pilot) a major part of the research energy of the unit was devoted to efforts to produce a satisfactory series of scales of flying skill. In these, certain standard maneuvers were specified, to be flown in a defined sequence. A score card was prepared which indicated a number of aspects of the maneuver which were to be observed and scored by the check pilot. Thus, on a steep turn the observer might have to score the angle of bank, time to complete the turn, and change in altitude. On a landing the observer might score part of field landed in, amount of bounce in the landing and attitude of the plane at the time of landing. The effort was to make all observations as simple and as quantitative as possible, and thus to have them be in terms of feet of altitude, degrees of heading, miles per hour of airspeed, and the like. These were supplemented where necessary, by more qualitative judgments of coordination of controls, amount of bounce, and so forth.

In flexibility gunnery, a great number of phase checks were developed by psychological personnel of the Research Division, Central School for Flexible Gunnery. One of these, for example, checked performance in stripping and assembling the .50 caliber machine gun. The task was broken down into the sequence of component operations. A score sheet was prepared listing each step in the sequence. The observer checked the student step by step on the score sheet, indicating by a simple check mark whether he did or did not perform each required step adequately and at the proper point in the sequence.

This type of check is well adapted to tasks for which a standard sequence is required and for tasks which are readily analyzed into a number of component elements. When the routines are less rigidly specified or the operations are more complexly integrated, the checking procedure becomes more difficult and would appear

to be less satisfactory. However, a number of check procedures were developed for quite complex sequences of tasks. Thus, the Psychological Research Project (Bombardier) developed a flight check to cover the complete course of a practice bombing mission, including preflight checks and all flight operations up to and including the actual simulated bombing of the target. A similar check was developed by the Psychological Research Project (Radar). Evidence on the objectivity of application of complex flight checks such as these is very meager. It seems clear, however, that they require a checker with a high level of experience and competence in the job duties for their effective use, and that some special training in the use of the check is indicated.

In potential relevance to the ultimate job criterion, subjectively scored job samples such as those which have been described would appear to be second only to the direct objective record of performance. Their rational basis is somewhat less satisfactory insofar as the behavior of the subject is mediated by an observer. However, the use of an observer permits a flexibility and scope considerably greater than that possible for an objective record. The possibilities of bias vary greatly for different instruments within this category, depending upon the type of observation which is required of the observer. Insofar as the observer serves as a simple recorder of instrument readings and of simple and precisely defined behavior items, bias is minimized. Insofar as more interpretation by the observer is required or permitted, more variation from observer to observer, time to time, and place to place is possible.

In these measures, reliability remains a problem. The disturbing factors are much the same as those discussed in connection with performance measures. However, insofar as the observer takes a significant role in the evaluation, the factors are somewhat changed. On the credit side, the introduction of the observer permits some allowance for variation in the objective external conditions. Thus, it is possible for an observer to make some allowance for visibility, turbulence, and the like, in evaluating a particular performance. Another advantage is the increased flexibility of a situation which includes the observer. This permits broadening the base of the observations, and the additional types of data which are included may be expected to make a contribution to the reliability of the total score. On the debit side are, of course, the fluctuations of the observer from moment to moment and variations among observers in standards of evaluation. Development of checks of this type requires that the scope of the evaluation be increased as much as possible, while at the same time the testing situation be so completely defined as to minimize variations from observer to observer.

Data on reliabilities of subjectively scored job samples are somewhat limited, but certain illustrative figures can be cited. For a series of elementary pilot training maneuvers, a median day-to-day retest reliability of .08 was obtained for 18 separate single items of performance scored on successive days by different checkers. On a different specific group of items, the reliability of a total scale of 16 items, selected from a total of 24 in terms of ability to discriminate between men with 15 and 55 hours of training, was .50 for 41 men with 55 hours of training and .39 for 35 men with 15 hours of training. A complete scale for basic instrument flying gave a test-retest reliability of .43 for 55 cases. This was, again, for retest on subsequent days with different check pilots. Data on the reliability of phase and flight checks are too limited to provide any empirical basis for evaluation of these techniques. They probably vary widely in reliability, depending upon the variety and simplicity of the behavior to be observed.

Rated Job Samples

Personnel of the Aviation Psychology Program were keenly aware of the difficulties and limitations which have filled the history of the use of rating scales. The tendency was, therefore, to resort to rating procedures only when no more objective procedure for scoring the details of a performance appeared to be available. However, there were a certain number of cases which were not successfully analyzed into behaviors which could be checked or scored in terms of specific performance, and in those cases rating procedures were used. The ratings varied widely in scope. At one extreme, ratings of a single maneuver or aspect of a maneuver were continuous with the type of scoring or checking referred to in the preceding section. At the other extreme were ratings of a complete mission or segment of training. Summary ratings of performance were also studied and these will be discussed later in connection with the general topic of summary evaluations.

Though rating procedures were developed by aviation psychologists only as a last resort, there were also many rating procedures already in effect as the standard procedures for evaluating performance in training. Typical of these was the check flight used as the fundamental basis for evaluation of proficiency in pilot training. In this evaluation, the student flew with a check pilot. He went through a series of maneuvers appropriate for his level of training, the particular choice and sequence of maneuvers being determined by the check pilot. After a flight of varying duration, for which an hour might be a roughly representative figure, the check pilot recorded a grade for the flight, together with such comments as he considered appropriate. The final grade represented a complex clinical evaluation of the perform-

ance of the student during that flight, together with whatever other factors of previous acquaintance with the student or the student's record might have influenced the observer.

The subjectivity of the above type of procedure and its extreme dependence upon the standards and judgment of the observer are obvious. Some degree of standardization may be achieved by centralized training of instructors and check pilots, Standardization Boards which review the ratings of individual check pilots, and the like, but at the best individual standards may be expected to show significant variation. Without strenuous efforts at standardization, variation from observer to observer is likely to become enormous. As a compensating advantage mention may be made of the fact that this procedure requires a synthetic judgment. There *may* be aspects of flight performance which are lost in an analytical approach, and it may be that a scoring of elementary performances can never give an entirely adequate evaluation of the over-all quality of pilot performance. If that is true, the synthetic rating has advantages. An additional advantage can be argued in that the rater is able to allow for the external conditions under which the flight was made, something that can hardly be achieved in more objective methods.

Rating procedures developed by aviation psychologists differed in many respects from those ratings which were routinely in use for the evaluation of proficiency. The differences centered around an effort to make the ratings more analytical and to have them reported in terms of described standards of behavior rather than on a scale of numbered or lettered steps. Analytical ratings on features of behavior were developed in hopes of relating the ratings more specifically to observable items of behavior rather than depending upon generalized impressions of the man. It must be admitted, however, that this attempt was generally only partially successful and that halo effects appeared generally to persist. The use of scale points described in some detail represented an attempt to improve consistency of interpretation from rater to rater and reduce variation in subjective standards. There is some evidence that this was achieved at least in part, for group to group variation in these ratings appeared to be less extreme, in several cases, than for uncontrolled letter or number grades.

In general, it is felt that rating procedures were inferior to those previously discussed in relevance to the ultimate criterion and especially in freedom from bias. The lowered relevancy arises from the fact that a stage of interpretation intervenes between what the man did and the score which he made. The variability in this interpretation concerns us when we consider reliability and bias; for the moment, our concern is that the interpretation

imposes *one more step* between the performance and the ultimate criterion. Though we may agree that how well a plane commander maintains the morale of his crew, for example, is an index of how competent he will be in his combat duties, we may be less willing to grant that how well he appears to an observer to motivate his crew is such an index. The additional stage of interpretation seems inevitably to weaken the rationale for the evaluation procedure.

It is in the matter of bias in particular that rating procedures appear to be weak. In these procedures, to a very large extent each rater provides his own standard. This standard will vary from rater to rater, from time to time, and from place to place. Comparison of groups in different schools or classes becomes meaningless, so that large-scale and long-time studies become impossible. Even when systematic procedures are introduced to assign to each rater members of each of the experimental groups which are being compared, bias is still possible. If the raters are aware of the group to which a subject belongs and if they are prejudiced in favor of some one of the particular training programs that are being compared, it is entirely possible for the rating of a man to reflect the bias towards the group of which he is a member. Therefore, especially in investigation of training procedures and in any investigation in which differences between classes or schools are of critical importance, rating procedures must be viewed with critical suspicion.

Adequate data on reliability of ratings are difficult to obtain. It is difficult to guarantee that ratings obtained from different individuals at the same station will be truly independent. At the worst, the presumably independent raters may cooperate directly in preparing the ratings. At the best it must be expected that both raters will be effected to some degree by the general reputation which is attached to the man at the particular station. Only intimate knowledge of the situation in a particular station will indicate how serious the contamination of separate ratings is likely to be. In any event, reliability coefficients for ratings will usually indicate consistency of a man's reputation at a given time and place rather than agreement based upon entirely independent observations of his behavior. Bearing this in mind, it may be stated that a number of the rating procedures *appear* to yield moderately satisfactory reliabilities. One study[1] showed the reliability of the summed check rides in primary pilot training to be about .80. A number of descriptive rating scales in operational training[2] gave correlations between two separate raters

[1] This study was reported by Lieutenant Robert J. Keller, then stationed at Psychological Research Unit No. 2.

[2] For discussion of these results see Report No. 16.

averaging about .50 or .60. However, the final interpretation of these ratings must remain in question. Data from correlations of summary evaluations suggest that the consistency of rating from one station to another is much less.

SUMMARY EVALUATIONS

We turn now to a consideration of summary evaluations of a whole period of training or of operational duty. These vary in detail, depending upon the degree to which they are based upon specific evaluations, the types of specific evaluations upon which they are based, and the manner in which the specific evaluations are compounded. At one extreme the summary may include nothing which has not already been recorded as a specific evaluation of a defined segment of behavior. Average circular error in bombardier training was such a summary evaluation, in that it represented a simple averaging of bombing errors on a specified series of training missions. At the other extreme, the summary may make no direct reference to any previous specific evaluation of behavior. One suspects that some over-all ratings, such as those of "officer quality" were of this type. The summary evaluations may involve in different degrees printed tests, objective performance scores, subjectively scored job samples, and rated job samples. The qualities of the final evaluation will stem in part from the qualities of these component elements. Finally, the summary evaluation may be a direct statistical compounding of the component scores, or it may represent a synthetic clinical judgment based upon them in unspecified ways and to an unspecified degree.

The conditions for a satisfactory summary evaluation would appear to be that it be in large measure based upon previous specific evaluations, that the specific evaluations themselves have desirable attributes as outlined in the previous section of this chapter, and that the procedures for combining the specific evaluations be objective and well-defined.

It seems unlikely that a summary evaluation will be of much value unless it is based upon previous observation of performance in specific situations. General after-the-fact impressions are notoriously untrustworthy and biased by irrelevant factors of general appearance, manner, and personal likeableness. An illustration of this was provided by certain ratings which were obtained of airplane commanders in operational training. At the same time that raters evaluated a group of men whom they had been instructing upon approximately 10 traits, they indicated what they considered to be the relative importance of each of the traits for over-all effectiveness in the job assignment. Though

"likeableness" was consistently placed at the very bottom of the list in importance, it nevertheless fell at the top in terms of its correlation with an overall rating. Though the raters disclaimed its importance, it still provided the chief basis for their over-all evaluation. The guarantee that specific evaluations have been made in advance is that they have been required to be officially recorded.

It seems obvious that the more relevant, accurate and unbiased the specific observations have been, the more relevant, accurate and unbiased will be the summary which can be extracted from them. Finally, the values of the component evaluations can only be maintained if they are objectively combined. The possibility of allowing for biasing external conditions, which is gained when records are clinically evaluated, would seem to be a small recompense for the introduction of the unreliability and personal bias of subjective interpretation into the final summary evaluation.

Summary evaluations, though differing in detail, seemed to fall in most cases into one of four categories. The categories were summary performance records, summary academic grades, summary ratings, and administrative actions. At this point some consideration will be given to each of these categories.

Summary Performance Records

In a number of cases, systematic provision was made for the recording and cumulation of objective performance records. These records are exemplified by average circular error for bombardiers, percent of hits in fixed gunnery for fighter pilots, air-to-air target firing or gun camera scores for flexible gunners, and the like. These records have all the appeal of relevancy to the ultimate task and freedom from bias which characterize the specific observations of which they are composed. They also present in an even exaggerated form the problems of control of external conditions which were discussed in connection with the type of specific evaluation of which they are composed. The problem of control of extraneous sources of variance, and consequently of attaining some minimum standard of reliability, is exaggerated in this case by the fact that the data are obtained under ordinary training conditions rather than under the conditions of a special experiment. This reduces the control of extraneous factors from that which *can* be obtained for purposes of research to that which typically *is* obtained in the normal course of training or operations. All the factors of weather, equipment, other personnel, character of the target and the like run rampant. The question becomes whether under these circumstances it will be possible to demonstrate any residual reliability associated with the persons or procedures being studied. It may be reiterated in passing that the

reliability need not be high to permit valuable research making use of the criterion, but it must be present.

Certain summary performance records appeared to be reasonably satisfactory from the point of view of reliability. For example, in a study at the Psychological Research Project (Pilot) the reliability of a series of air-to-air fixed gunnery missions amounting to 1200 rounds of firing was estimated as .63 (N = 1064). On approximately the same group, the reliability coefficient for 400 rounds of air-to-ground gunnery was .59. In other cases, the reliability appears to be much less satisfactory. A number of of estimates of between-missions reliability are available for circular error in bombardier training at the Training Command level.[3] They give the following results:

Class	N	Reliability	
41-C	70	.27	
43-3	129	.03	
43-1,2,3,4	94	.06	Same population, different grouping of scores.
	128	.37	
	100	.16	
	63	.06	
	172	.13	
	68	.09	
	174	.01	
	102	.08	
645	94	.05	Same population, different grouping of scores.
	94	—.08	
	94	.16	
	94	—.02	
	94	.18	
	89	.18	

The median of all these separate values is .08. This provides rather a crude estimate of reliability but it does not provide very strong assurance for the use of this criterion as an evaluation of individual proficiency. As applied to *crew proficiency*, where pilot, navigator, bombardier and enlisted crew members remain together, the reliability appears to be somewhat higher.[4] In this case, personnel in the plane is held constant. The higher reliabilities confirm other findings which indicated that circular error was as much a function of the pilot as of the bombardier.

As an objective evaluation of the proficiency of radar observer performance, results were available on circular error in radar bombing. Analysis of available data gave the following results for odd vs. even missions:

	r	*N*
Boca Raton, av. of 3.2 missions	.32	112
Victorville, av. of 4.3 missions	.20	372

The results for these different criteria in different types of training illustrate the range of reliabilities which were obtained. It is clear that some summary performance records may be quite acceptable, while others appear quite unsatisfactory in this regard.

[3] See Report No. 9 of this series for details.
[4] See Report No. 16 for further data.

Summary Academic Grades

Though used to some extent by aviation psychologists, academic grades appeared less clearly relevant to the ultimate criterion than many other types of criteria. In some cases, where the job appeared to involve substantial intellectual components, as in the case of navigator and flight engineer, the rationale for accepting academic grades as relevant criterion measures seemed somewhat better, and in these cases some use was made of that type of criterion. Particularly in the case of pilot, with its emphasis upon performance and skill, little attention was paid to measures of academic proficiency in ground school courses.

The chief drawback in the case of academic grades appeared to lie in the lower level of relevancy. Though many of the specific evaluations of academic performance used in routine training lacked technical polish and suffered from subjectivity of evaluation, the summary evaluations *did* ordinarily come in a fairly direct and explicit fashion from actual specific evaluations. That is, there were actual tests, recitations, and work samples underlying the grade, and it was ordinarily compounded from specific scores and ratings of this sort in a uniform and objective manner. Available evidence indicates that most such grades were moderately reliable. For example, in navigation training the reliability of examination grades was estimated as .90 in one class of about 300, the reliability of classroom grades .82, and the reliability of flight grades .72. These are based on the correlation of odd with even weeks, and insofar as the grading was subjective some spurious relationship may be present. However grades did lack a stable reference point, so that freedom from variation from time to time and place to place cannot be claimed for them. Grades were subject to bias depending upon the standards of the station at that time, and more particularly the standards of the specific instructor or group of instructors. In those types of training in which grades were studied, appreciable variation from station to station and from flight to flight within a station was uniformly found.

Summary Ratings

A great variety of summary ratings were in use in the routine evaluation of aircrew personnel. These included routine efficiency ratings, required to be submitted on all officer personnel; ratings on officer qualities of cadets in training, used to determine whether the cadet in question should be commissioned a 2nd lieutenant or appointed a flight officer; ratings of pilots at each stage of training on flying skill, maintained as a cumulative record for each man; and a great variety of ratings of other specific groups for specific purposes. Most of these ratings were on a simple scale

of 3, 5, or more numbered or lettered points, which may have been further identified by such brief general descriptive labels as "superior," "above average," or the like. A certain number of additional scales were developed by aviation psychologists for research purposes, where no better procedure seemed readily available. These took the form of descriptive scales, in many cases, defining certain traits to be rated and describing degree of possession of the trait. However, the limitations of these procedures were keenly felt, and no great confidence was placed in them as research tools.

The limitations of rating procedures as applied to rating a specific segment of behavior have been described in a previous section. All these are present in summary ratings and others as well. It is an unfortunate characteristic of summary ratings that they are frequently not based in any clear way upon previous evaluations of specific behavior. The limitation may involve either the amount of specific information, the technique for synthesizing it, or both. In the extreme case, which is only too close to reality, a summary rating represents an over-all judgment of an individual, rendered after a longer or shorter period of experience, given with no basis of previous systematic observation and evaluation of the individual. There are often no data to refer to in the form of flight checks, tests, or performance records. The rating represents merely the unguided, intuitive impression of the rater. In this case, the rating will obviously reflect personal bias, and insofar as no other data are available it may be expected to reflect nothing else but personal bias. Its freedom from bias will be low, and since biases are likely to be individual and are almost certainly unrelated to the ultimate criterion, the rating is also likely to have little to recommend it on the score of reliability or of relevance. An appearance of reliability may arise due to the general reputation factor which was discussed in connection with specific ratings. It may be anticipated, however, that this will not hold up except within a limited group. It may be stated in passing that very little success was ever achieved in the Aviation Psychology Program in predicting ratings of this kind.

Not all ratings are as bad as the type we have just described. In some cases, a summary evaluation in the form of a rating may be based upon a reasonably extensive set of explicit specific evaluations, which were made and recorded as training progressed. In some cases, day to day evaluations may have been implicit in the relationship between the rater and the persons rated. Even in these cases, however, the use of a clinically based rating as the technique for summarizing the earlier evaluations introduces an element of subjectivity and bias into the final result which can hardly fail to prejudice its value as a criterion.

Administrative Actions

There were a number of administrative actions which were taken with regard to aircrew personnel which provided summary evaluations of proficiency and presented possibilities as criterion data. Logically, these are closely akin to the ratings which have just been discussed, but in terms of their practical importance and of consequent possible differences in the manner in which they were prepared, they appear to merit separate consideration.

The administrative decision which served most often as a research criterion in the Aviation Psychology Program was the decision to graduate (or to eliminate) a man from a particular phase of training. Elimination because of lack of proficiency or for reason of fear or at own request provided a readily available criterion of proficiency which appeared to have some relevance both from the positive and the negative point of view. On the one hand, the skills and techniques which had to be learned in training provided the foundations for operations in combat. It seemed rational to believe that those who were particularly apt in learning the basic knowledges and skills would, in general, be those who would be proficient in later stages of operations. That is, training performance generally appeared to have some relevance for the ultimate criterion of combat performance. On the other hand, it appeared to be important to select for training those individuals who would in fact complete and be graduated from training, and thus be available for assignment to combat duty. It may be argued that those who were eliminated from training who *could* have become successful in combat should never have been eliminated in training, and that procedures of training and training eliminations were at fault and should have been changed. In the long run this is true. But working within practical limitations of time and an existing training situation, it may still be important to pick men who will succeed in that training situation. That is, training performance appears to have some direct relevance for its own sake.

Other administrative actions which were studied, and to some extent used as criteria, included reevaluation and removal from flying by Flying Evaluation Boards, promotions, decorations, assignment to first pilot vs. co-pilot duty, assignment to lead crew, removal from combat operation because of operational fatigue, and the like.

Practically all administrative actions imply a rating. They differ from many other ratings, however, in the practical importance of the rating which is made. Something is clearly going to be done on the basis of the rating. A man will be eliminated from training, removed from flying status, put in a position of critical importance and the like. On the basis of this, we may expect

that the evaluation will be more thoughtfully and conscientiously made than will be the case when the rating is merely an administrative chore. Relevant records will be consulted, testimony will be assembled and weighed, and the worst qualities of ratings somewhat mitigated. It must be recognized, however, that most administrative actions do fundamentally imply ratings, and that the limitations of rating procedures inhere in them.

CHAPTER FIVE

Determining the Validity of Single Tests

In this chapter consideration will be given to the statistical procedures which were used in computing indices of validity for single classification tests. The next chapter will then consider problems concerned with combining a group of tests into the most effective test battery. A description of the typical computational procedures will first be given. Then certain special problems which arose, for which solutions were reached in part, will be presented.

COMPUTATIONAL ROUTINES

The validity of a single test for predicting a particular aircrew criterion was uniformly expressed in terms of a coefficient of correlation. In the case of criteria which provided a continuous distribution of criterion scores, for example, bombing circular error, percent hits in aerial gunnery, etc., product moment correlation coefficients were computed. For these, as for other statistics, work sheets were developed to facilitate procedures of computation. The computation procedures at different units differed somewhat, depending upon the previous training of the personnel responsible for statistical work at the unit in question. To guarantee efficient computational procedures and adequate checks, a correlation chart was finally issued by the Psychological Section, Headquarters AAF Training Command.[1] A discussion of this form and its use is presented in the Appendix.

Many criteria provided only a dichotomous division of the group being studied into such categories as graduates and eliminees. With these, the alternative was between computing a biserial correlation coefficient or a point-biserial. The formulas for these are respectively:

$$r_{bis} = \frac{M_1 - M_2}{S.D._t} \cdot \frac{pq}{z}$$

and

$$r_{pbis} = \frac{M_1 - M_2}{S.D._t} \sqrt{pq}$$

[1] Lt. Col. Philip H. Dubois was primarily responsible for developing this form.

703315—47—5

Since the formulas differ only by the factor $\sqrt{\frac{pq}{z}}$, which is a constant for all correlations computed from a given group, for a set of correlations based on a single group, the relative sizes of the validity coefficients of different tests (and consequently of their multiple regression weights) are the same for the two types of coefficient. Within a single sample, then, it makes no practical difference which type of coefficient is used. Practical issues arise, however, when it is necessary to combine data from several samples in which the proportion in the "graduate" group differs, or when it is necessary to correct an obtained correlation coefficient for restriction of range due to selection of the group sent into training.

In practice, biserial correlation cofficients rather than point biserials were computed in most cases in the Aviation Psychology Program. The derivation of the biserial correlation coefficient assumes that the dichotomized variable is basically continuous and normally distributed. The dichotomy is considered to be arbitrarily imposed by some administrative condition and not to represent any general or necessary break of the group at that particular point. The biserial coefficient has the advantage that when the above conditions are satisfied, the value obtained for the correlation coefficient is independent of the point at which the group is split. This is not true for the point biserial, which will be larger if the group is split into nearly equal sub-groups than it will be if the split is made into one large and one small group.

Since elimination rates in a given type of training, to consider the most frequently used type of dichotomous criterion, varied markedly between schools, Commands and classes, the variation in value for the point biserial was a matter of very real concern. Thus a biserial correlation coefficient of .50 against pass-fail in primary pilot training would have corresponded to a point biserial of .39 in class 43-G, in which the elimination rate was approximately 38 percent, but would have corresponded to a point biserial of .31 in class 44-E in which the elimination rate was approximately 12 percent. That is, the same basic relationship would have given values differing by about 25 percent if the point biserial had been used in these two cases. The difference is clearly quite an appreciable one. Though the example cited represents the extreme deviation for complete primary pilot classes, differences between single schools, as well as for other types of training, were frequently as large as this. The resulting effect upon validity coefficients becomes, then, of practical as well as theoretical significance.

The comparability of values obtained from several samples in which the percentage graduating varies, which permits the direct combination of the results, led to the choice of the biserial coefficient of correlation for routine use. However, we should examine somewhat further the assumptions underlying that coefficient.

The assumption that the variable underlying graduation-elimination is continuous seems entirely reasonable. This was borne out by a wide range of graduation rates in different schools, classes, or Commands, even when the quality of the entering population did not differ. It is fairly clear that the particular point at which the division between graduation and elimination was made depended upon conditions which were local and temporary. The assumption of a normal distribution introduces more serious problems, particularly when an appreciable proportion of applicants for training have been disqualified because of low aptitude test scores. It is possible, and perhaps reasonable, to consider that the skill in question would have been normally distributed *either* in the total group of applicants for a given type of training *or* in the fraction selected for training on the basis of some type of screening procedure, but if the screening had any validity at all, the distribution of skill could not have been normal in both cases. The more reasonable assumption would probably be that it was normal in the unrestricted population. If this is the case, the biserial correlation coefficient was not strictly applicable in the case of curtailed groups. This fact makes many of the validation statistics, especially for the later classes, which were more sharply curtailed, somewhat in error, though whether the error introduced by the assumption of normality is a serious one has not been determined. This point will be considered further when the problem of correcting for restriction of range is considered.

For some variables which were available for study as prediction measures, the variable itself was dichotomous or fell readily into dichotomous form. These were variables such as marital status, first preference regarding type of training, presence or absence of previous flight training, and the like. In these cases, three types of coefficients are possible, depending upon what assumption is made as to the continuity or non-continuity of the basic distributions. If both distributions are assumed to be continuous, and both dichotomies to be artificial, the tetrachoric correlation coefficient gives an estimate of the product-moment correlation in the normal frequency surface which has been cut by the two dichotomies. These were determined by using Thurstone's computing diagrams.[2]

[2] Thurstone, L. L., et al, Computing Diagrams for the Tetrachoric Correlation Coefficient. Chicago: Univ. of Chicago Bookstore, 1933.

If one dichotomy is real and the other artificial, a coefficient analogous to the biserial correlation coefficient may be computed, treating the real dichotomy as a point score. This may be called ϕ_{bis}. It can be obtained from the formula for the biserial correlation coefficient reported in an earlier section by assigning point scores such as 1 and 0 to the two categories of the real dichotomy. The formula simplifies in this case to

$$\phi_{bis} = \frac{ad - bc}{z\sqrt{pq}}$$

where a, b, c, d, = the entries in the 4 cells of the fourfold table
p and q = the percents in the two categories of the artificial dichotomy, and
z = the ordinate corresponding to the values p and q.

If both dichotomies are necessary and genuine dichotomies, as for example, in the correlation between marital status and first choice for (or not for) pilot training, then the relationship between the two variables may be represented by the phi coefficient, a coefficient in which each variable is treated as having a point distribution, and the correlation is for a fourfold point-surface. The formula becomes

$$\phi = \frac{ad - bc}{\sqrt{pqp'q'}}$$

where the meaning of a, b, c and d are as above and p, q, p′ and q′ refer to the percents in each category of each of the dichotomies.

When a table of correlations combines data from continuous and dichotomous variables, some question occasionally arises as to which of the above is the appropriate coefficient to use in combination with product moment correlations among the continuous variables. We have already discussed the case in which the criterion is dichotomous. We must now consider the case in which one of the prediction variables is dichotomous. When the dichotomy in the prediction variable is a natural one or *whenever the practical data to be used will be gathered in such a way that they must be used as a dichotomy,* the appropriate coefficients to be used are as follows:

(a) For correlations with continuous variables, the point biserial correlation coefficient.

(b) For correlation with another dichotomy which is considered to be an artificial dichotomy (i. e., where biserial correlations are used between continuous variables and the dichotomy), the biserial phi coefficient (ϕ_{bis}).

(c) For correlation with another natural dichotomy (i. e., where point biserial correlations are used between continuous variables and the dichotomy), the phi coefficient (ϕ).

When the dichotomy in the prediction variable is an artificial one *and the variable will be available for use as a continuous variable when it comes time to make practical use of it,* the values to be used in the three cases above are as follows:

(a) For correlation with continuous variables, the biserial correlation coefficient.

(b) For correlation with another artificial dichotomy, the tetrachoric correlation coefficient.

(c) For correlation with a natural dichotomy, the biserial phi coefficient.

ITEM VALIDATION IN TEST CONSTRUCTION

In the course of the test development program for the Air Forces aircrew classification tests, two rather different types of tests were developed. One type of test was designed to be a measure of a relatively homogeneous function. Examples of this type were tests of numerical operations, reading comprehension, figure analogies and the like. In this type of test, preliminary internal consistency item analysis was ordinarily carried out in order to increase the homogeneity of the test materials. In revising the preliminary form, those items were retained which showed satisfactory correlation with total test score.

The second type of test was made up of more or less heterogenous items. Typical of these tests were a Biographical Data Blank covering various items of personal information, a Sports and Hobbies Information Test, a test of satisfactions, and various temperament and personality questionnaires. For this type of material, item analysis in terms of internal consistency is essentially meaningless because no effort is being made to get a pure test of a single homogeneous function. This type of test puts together a group of items which are related only very loosely in terms of the kind of question which they ask or the label under which they may be grouped. In these instances it becomes not only appropriate but necessary to evaluate each of the separate test items in terms of its contribution to the validity of the total test score and even of the total testing battery. In a perfectly real sense, each item in these cases may be thought of as a separate test. It becomes necessary, therefore, to validate each item for its own sake. Given that time and personnel were of no concern and the available population was sufficiently large to provide stable values, it would be appropriate to determine a validity coefficient for each item, to determine all the item inter-

correlations, and to determine a regression weight for each of the separate items. Ordinarily, practical conditions will not justify such a detailed analysis of the separate items, which may total as many as several hundred.

The procedure most typically used in the Aviation Psychology Program was to obtain an indication of the validity of each separate item by comparing the percent of successful individuals responding to that item in a particular way with the percent of unsuccessful individuals responding in that same way. When such percents had been determined, it was possible to translate them into a tetrachoric correlation or phi coefficient. Evaluation of each item was based upon its validity, without regard to its correlations with other items or tests. This procedure was adopted not because it was believed to be the most adequate and most elegant one but because it represented a practical undertaking from the point of view of time and effort involved. Furthermore, more elaborate statistical procedures involving correlations of items with each other and with tests were not used because certain of the problems involved in making appropriate use of more complex item data had not been solved analytically.

Whatever index of individual item validity was involved, the next step was to prepare a scoring key including those items in the test which showed the best individual item validity. Since it is well known that a group of keyed items will have a somewhat lower validity on a second sample, due to sampling error in the original determination of the individual item validities, the crucial question is what the validity of this group of keyed items will be on a new sample. The approach to the problem which was most extensively used in the Aviation Psychology Program was to break the original sample into two parts, run two separate original item analyses, and then carry out a cross-validation study to determine the validity of the score based on items selected from one half as applied to the other half. Another approach, which is much more profligate of time, would be to administer the test to a new sample scoring those items selected on the basis of the initial validation and then wait for criterion data to mature in order to determine the validity of the selected items on the new sample.

Neither of the above methods is entirely satisfactory. In neither case do they provide any analytical procedure for determining how many items should be included in the scoring key. Obviously, one starts by including the most valid items, but the problem is how far down the list one should go. This problem is complicated by the fact that validities obtained in a single sample will in general regress in a new sample. At a certain point, the addition of more items having some slight validity in a particular sample

will have the effect of introducing an undue proportion of nonvalid variance and attenuating the validity of the total key. In the second place, neither of these methods makes use of all existing data. At any given time, the scoring key which will have the highest validity for a new sample will be the one which is based upon the largest possible number of cases. However, if all cases are used to determine the scoring key, none are left to provide an unbiased estimate of validity in a new sample so that satisfactory data will be available for the empirical determination of the weight which should be given the score in arriving at a composite score.

It has been indicated that item validation was carried out primarily for heterogeneous test materials. In proportion as the materials appeared homogeneous, little need was felt for the validation of single items. In a completely homogeneous test, item validity is by definition a direct function of item internal consistency. Of course, the test which approximates this condition in practice is probably rare, and item validation was in fact carried out for a number of tests which were designed to be homogeneous. Before giving much weight to such analyses, one should first have some assurance that the spread of item validities is in fact greater than would be expected by sampling alone. If differences in item validity represent only the sampling fluctuations among truly homogeneous material, the differences in item validity may be expected to disappear in a new sample, and the labor of selecting items in terms of their individual validities will have been entirely in vain.

PROBLEMS OF RESTRICTION OF RANGE

The research program for development of aircrew classification tests brought into sharp prominence certain statistical problems which have long been recognized and for some of which partial statistical solutions have long been available. These are the general problems of inferring statistical parameters in a population from those which have been obtained in a sample when the sample has been curtailed in some way with respect to the range of one or more variables. This problem was particularly acute in the aircrew classification program because a number of selective procedures operated at successive stages of classification or of training. The samples upon which criterion data became available had frequently been sharply restricted in some way as compared with the population tested with classification tests.

A number of different types of curtailment operated in the classification program. The most frequently occurring situation with regard to curtailment was that in which a population was tested with a group of classification tests and the men to be

assigned to a particular type of training were then selected at least in part upon the basis of performance on a weighted composite of the classification tests. This selection was in part negative in the sense that minimum qualifying scores were established for the weighted composite (called "stanine" in the aircrew testing program) below which applicants were not accepted for that type of training; it was in part positive, in that an effort was made to send men with the highest stanines for a particular aircrew specialty into that type of training. When only a select sample was sent into a particular type of aircrew training, it was necessary to correct the validity coefficients obtained for that sample in order to have an unbiased estimate of the validity coefficients in the total population tested with the classification tests. This was necessary because weights for predicting aircrew success were to be used with the total population tested rather than merely with that fraction of it assigned to a particular type of training.

A more complex type of curtailment arose when data with respect to the AAF Qualifying Examination were being analyzed. In this instance the men for whom data on actual success in training became available had first been screened by requiring a minimum qualifying score on the Qualifying Examination and subsequently by specifying minimum qualifying scores and by some degree of positive selection upon the aircrew stanine in terms of which they were finally assigned to training. Since the population for which validity estimates were desired was in this instance the complete population to which the AAF Qualifying Examination had been administered, it was appropriate to correct both for curtailment on the Qualifying Examination and for subsequent curtailment on the aircrew stanine.

A second more complex type of situation was involved when it was desired to use data on success at an advanced stage of training or at the level of combat to provide an estimate of the relative validity of the different classification tests for predicting that type of performance. In this instance it was again legitimate to ask what the validity of the tests would have been if all men tested had reached that stage of performance, because the decision as to which men should be accepted for training had to be made in the case of every man tested. We are interested in knowing how well the men who are disqualified either by tests or by earlier stages of training would have done if they had been permitted to continue to the level of training or operations currently under study.

Formulas are available to correct correlation coefficients for the effect of restriction of range, provided the data conform to certain conditions and providing that certain necessary statistics

are available for the group in question. A solution of the problem in the case in which all the variables being studied are normally distributed was devised by Karl Pearson in 1903.[3] Formulas were derived for the several possible cases of curtailment on a single variable, and a generalized solution was presented for the case of curtailment on more than one variable. The several cases of the single variable and the generalized formula are discussed below.

Using the notation

S_i = SD of variable i in the unrestricted distribution
s_i = SD of variable i in the restricted distribution
R_{ij} = Correlation between variables i and j, in the unrestricted distribution
r_{ij} = Correlation between variables i and j in the population which has been directly or indirectly restricted,

it is possible to arrive at the formulas for three distinct cases of restriction on a single variable.

a. *Case 1.* When the restriction is in variable *1* and the ratio of the two standard deviations of variable 2 is known:

$$R_{12} = \sqrt{1 - \frac{s_2^2}{S_2^2}(1 - r_{12}^2)} \qquad (1)$$

Example: Formula (1) would be used in estimating the correlation between a research test and pilot stanine, when the distribution has been restricted on the basis of pilot stanine, and the ratio of the standard deviations of the restricted and unrestricted ranges of the test is known. Stanine would be variable *1*; test scores would be variable *2*. This situation was rarely encountered in practice.

b. *Case 2.* When the restriction is in variable 1, and the ratio of the two standard deviations of variable 1 is known:

$$R_{12} = \frac{r_{12}\frac{S_1}{s_1}}{\sqrt{1 - r_{12}^2 + r_{12}^2\frac{S_1^2}{s_1^2}}} \qquad (2)$$

Example: Formula (2) would be used in estimating the correlation between pilot stanine and graduation-elimination, when the distribution has been restricted on the basis of pilot stanine, and the ratio of the standard deviations of the restricted and the unrestricted distribution of pilot stanine is known. Stanine would be variable *1*; graduation-elimination would be variable *2*.

[3] Pearson, K. Mathematical contributions to the theory of evolution—XI. On the Influence of natural selection on the variability and correlation of organs., Phil. Trans. Royal Soc. of London, Series A. 200, 1903, pp. 1-66.

c. *Case 3.* When the restriction is in variable 3 and the ratio of the two standard deviations of variable 3 is known:

$$R_{12} = \frac{r_{12} + r_{13}r_{23}\left[\frac{S_3^2}{s_3^2} - 1\right]}{\sqrt{\left[1 + r_{13}^2\left(\frac{S_3^2}{s_3^2} - 1\right)\right]\left[1 + r_{23}^2\left(\frac{S_3^2}{s_3^2} - 1\right)\right]}} \quad (3)$$

In some instances, r_{13} is not known and formula (3) must be expressed in terms of R_{13}. Formula (3) then becomes:

$$R_{12} = \frac{r_{12}\sqrt{1 + R_{13}^2\left(\frac{s_3^2}{S_3^2} - 1\right)} + R_{13}r_{23}\left(\frac{S_3}{s_3} - \frac{s_3}{S_3}\right)}{\sqrt{1 + r_{23}^2\left(\frac{S_3^2}{s_3^2} - 1\right)}} \quad (3a)$$

Example: Formula (3) or (3a) would be used in estimating the correlation between a test and graduation-elimination, when the distribution has been restricted on the basis of pilot stanine, and the ratio of the standard deviations of the restricted and unrestricted distributions of pilot stanine is known. Test scores would be variable *1*; graduation-elimination would be variable *2*; stanine would be variable *3*. Formula (3) would be used if the test-stanine correlation is based on the restricted sample and formula (3a) if this correlation is based on the total population.

That the corrections for restriction of ranges were more than an academic matter may be seen by comparing the validity coefficients obtained from a complete group and from the fraction of that group which met the relatively high standards which were in effect for admission to pilot training at the end of the war. These particular data are based on the "experimental group," a group of men who were tested and then entered into pilot training without regard to their performance on the tests. Validities are presented both for the complete group tested and for that fraction of the group which both passed the AAF Qualifying Examination and achieved pilot stanines of 7. The results are as follows:

	Total Group ($N = 1036$)	*Qualified Group* ($N = 136$)
Pilot Stanine	.64	.18
Mechanical Principles	.44	.03
General Information	.46	.20
Complex Coordination	.40	—.03
Instrument Comprehension	.45	.27
Arithmetic Reasoning	.27	.18
Finger Dexterity	.18	.00

It can be seen that where the restriction is as severe as this, amounting to the exclusion of about 87 percent of the cases,

the changes in the resulting correlations are very striking. The small size of the "qualified" group makes the results somewhat unstable. However, it can be seen that the curtailment reduces stanine validity by over forty points, reduces the validity of tests weighted for pilot by about thirty points on the average, and reduces the validity of tests which are not specifically pilot tests by about fifteen points. These exact values are not particularly important because of the small number of cases involved. The chief point is that the shifts are decidedly large. Though during most of the work of the Aivation Psychology Program, the amount of selection was not as severe as in this example, the effect was sufficiently marked so that raw obtained correlations would often have been meaningless unless correction formulas were applied.

Pearson's article also included formulas for the general case, in which curtailment took place on more than a single variable. These formulas are quite involved and are not presented here. The same essential formulas for the general solution, but based on somewhat different assumptions and expressed in more convenient form have recently been reported by E. Reeve.[4] The basic assumptions of this derivation are:

(1) that the regressions of the nonselected variables on the selected may be treated as rectilinear throughout the total population, and

(2) that the variability of the nonselected variables is the same for each value of the selected variables.

The following notation is used:

x = any of the variables which is not directly restricted.

a = any of the variables which is directly restricted.

r = matrix of correlations in restricted group.

R = matrix of correlations in unrestricted population.

H = diagonal matrix giving ratios of standard deviations $\left(\frac{\Sigma}{\sigma}\right)$ of unrestricted to restricted group.

b = matrix of partial regression weights (beta weights) in restricted group.

In this case it can be shown, in matrix notation, that

$$R_{xa} = R_{aa}\, H_a b_{ax} H_x^{-1} \tag{4}$$

and

$$R_{xx} = H_x^{-1}\,(r_{xx} - b'_{xa}\, r_{ax} + b'_{xa} H_a\, R_{aa} H_a\, b_{ax})\, H_x^{-1} \tag{5}$$

where

$$H_x = 1 - b'_{xa} r_{ax} + b'_{xa} H_a\, R_{aa} H_a\, b_{ax} \tag{6}$$

The matrix notation used above presents a rather extended series

[4] Reeve, E. Correcting for Selection. Unpublished report supplied informally to Maj. Roger Russell, AAF. These formulas were also derived independently by Lt. Col. A. P. Horst of the Aviation Psychology Program.

of operations quite compactly, and has the additional advantage of suggesting layout of work sheets and order of operations for carrying out the necessary calculating procedures. It can be seen that the computations will be quite laborious at best when several variables are directly restricted.

The formulas which have just been discussed were all derived for product moment correlations as these are obtained for continuous variables. They are not strictly applicable to biserial correlations obtained from dichotomous variables. As has been stated previously, it is not possible for a variable both (a) to satisfy, in a restricted group, the requirements for applicability of the biserial correlation formulas and (b) to satisfy in the unrestricted population the conditions for use of the above formulas for correction for restriction.

In the simple case of direct curtailment on a single variable (Case 2 above), a technique for obtaining an estimate of the biserial in the unrestricted population from data available in the restricted sample was reported late in the war period by Gillman, and Goode.[5] This is essentially a procedure for obtaining a least-squares estimate of the slope of the regression line from the data on the part of the distribution which remains after truncation. The procedure is as follows:

Let G = correlation estimated from this procedure (subsequently referred to as a G-coefficient)
f = number of subjects with score in interval aVI xVI b
p = fraction of these falling in passing group
u = standard abscissa value corresponding to $p = p_u$

$$X = \frac{z_a - z_b}{p_a - p_b} \tag{7}$$

Then compute

$$N = \Sigma f,\ \Sigma fX,\ \Sigma fX^2,\ \Sigma fu \text{ and } \Sigma fXu$$

From these

$$A' = N\Sigma fXu - (\Sigma fX)(\Sigma fu) \tag{8}$$

$$D = \Sigma fX^2 - (\Sigma fX)^2 \tag{9}$$

Then

$$\tan \theta = -\frac{A'}{D} \tag{10}$$

and

$$G = \sin \theta \tag{11}$$

The computing procedures outlined above provide a technique for estimating the correlation in the population in the simplest case, in which the correlation is between the variable which has

[5] Gillman, L. and Goode, H. H. An estimate of the correlation coefficient of a bivariate normal population when *X* is truncated and *Y* is dichotomized. Harvard Educ. Rev., 16, 1946, 52-55.

been directly restricted and the dichotomous criterion. The assumption of normality in the unrestricted population is still involved, but no assumption need be made as to the nature of the distribution in the restricted sample of the variable which underlies the dichotomy. In this regard, therefore, the procedure which has just been described is much to be preferred to the procedures which require the computation of a biserial correlation coefficient in the curtailed group and its subsequent correction.

No procedures analogous to the one just described are known for the case of indirect curtailment (Case 3 above) or for curtailment on more than a single variable. Unfortunately, these were the situations which arose most frequently and most critically in the Aviation Psychology Program. Whenever a single test was being studied, rather than the stanine, Case 3 was involved, and whenever any advanced type of criterion was under study curtailment had taken place on several variables.

As indicated above, existing formulas for correcting for curtailment are not strictly applicable to biserial correlation coefficients. No analytical solution is available to indicate the direction and amount of the error which is involved when existing correction formulas are used in these cases. However, one set of artificial data was studied[6] to obtain empirical data upon the direction and extent of the errors involved. This was carried out only for the simplest case (Case 2 above), in which direct restriction upon a single variable is involved. Tables of synthetic data were prepared for a stanine validity of .50 and for various elimination rates. From these tables, curtailed groups were set up, eliminating first the 1's, then the 1's and 2's, etc. Biserial correlations were computed from these data and were corrected by formula 2 above. It appeared that:

(1) In these cases, which were designed so that the dichotomy in the restricted group was more uneven than in the unrestricted, the correction formula uniformly tended to underestimate the true value.

(2) The underestimation increased as the amount of curtailment increased and as the unevenness of the dichotomy in the unrestricted population increased. In the most extreme case studied, in which 60 percent were disqualified on the basis of stanine and in which the population split of graduates and eliminees was 90-10, the true value was underestimated by 20 percent.

These findings suggest that in the case of correction of stanine validities, the general tendency of the Pearson formula was to underestimate the true values, and that this underestimation was

[6] These analyses were carried out under the direction of Capt. Lloyd Humphreys.

most severe during periods of low over-all elimination. The underestimation should also have tended to become greater as the stanine requirements for admission to training were made more stringent. The indications from this analysis of artificial data are not borne out by empirical comparisons of values using the Pearson formula and using the G-coefficient discussed on page 68. Data were analyzed for 20 classes in primary pilot training, totaling about 137,000 men, 15 classes in advanced navigation training, totaling about 10,000 men, and 9 classes in bombardier training totaling about 7,000 men. Biserial correlations were computed and were corrected by the Pearson formula, and population values were also estimated using the G-coefficient. The median obtained biserial for the 20 pilot classes was .41, the median corrected biserial .52, and the median value for the G-coefficient .50. The corrected biserial was as much as .07 higher and as much as .06 lower than the G-coefficient in single pilot classes. For navigators, the median value for new aviation cadets was .43 for the uncorrected biserial, .61 for the corrected biserial and .59 for the G-coefficient. The range of differences between the two estimates of the population value was from +.07 to —.03. For bombardiers, the median value for new aviation cadets was .24 for the uncorrected biserial, .28 for the corrected biserial and a .32 for the G-coefficient. The range of differences between the two sets of population estimates was from .00 to —.15. The use of the G-ciefficient was rendered somewhat questionable in the case of the bombardiers, due to the fact that graduation rate was 100 percent for certain stanines in certain classes.

The lower values for the G-coefficient in the case of pilots and navigators, as compared with the Pearson formula, exactly reverse the situation found in the previous analyses of artificial data. Two possible explanations are offered. In the case of pilots, the distribution of augmented stanines is very far from normal, due to a piling up at stanine 9. This results from the addition of a special credit for flying experience in the case of some 10 percent of the men. The artificial convention was adopted that no stanine higher than 9 would be given. This produced a piling up, and in most cases a secondary mode, at stanine 9. Since both procedures assume a normal distribution, the lack of normality in the augmented pilot stanine may have distorted the expected relationships.

A second explanation may lie in the values which were assumed for the stanine standard deviation in the unrestricted population. The stanine score was originally set up so that each point on the score scale represented one-half of a standard deviation of the distribution of raw composite scores. Limiting the number of steps to 9 forced the extreme tails of the distribution into the 1

and 9 categories and reduced the scatter of the group so that the theoretical standard deviation was 1.96 rather than 2.00. However, empirical population values for certain periods and certain aircrew specialties fell distinctly below the theoretical values. In other words, either the population became less variable or the conversion tables were faulty. The situation was somewhat further complicated in the case of pilots by the matter of flying experience credit. The addition of this credit resulted in an increased standard deviation for the distribution of stanine scores. Empirical studies indicated the increase in standard deviation to be approximately 0.10, so that an assumed population value of 2.10 was used in correcting these values. Here again the assumed value appears often to have been somewhat larger than the value actually obtained for populations tested from month to month. The over-estimation of the population value would lead to an over-correction using the Pearson formula and would account for the obtained discrepancy between the two methods. In using the correction formulas the problem of whether to base population standard deviation estimates upon the theoretical stanine distribution, or upon empirical values for limited time periods was a troublesome one.

Two further general problems should be discussed in connection with the topic of restriction of range. In the first place, a basic assumption which must be made in any inference to a total population from data on a restricted group is that the criterion variable in the restricted group is not qualitatively different from what it would be in the population. That is, one must assume that within a restricted group elimination of the less apt students is made upon the same bases and with the same sharpness of discrimination that would be the case in the unrestricted population. There is at least some reason for doubting the correctness of this assumption. General observation of training programs and elimination procedures suggested that as only the more apt men were sent into training, while administrative pressure was kept up to hold to a specified, standard elimination rate, factors other than proficiency entered in increasingly to determine whether or not a given individual should be eliminated. This would tend to be true in any case if discriminations of degrees of ability are more difficult to make at the higher than at the lower ability levels. A final difficulty with corrected correlation coefficients was that standard error formulas for the corrected values were not available, so that it was not possible to establish the precision of the resulting estimated values. It seems probable that the standard errors will be substantially larger than those for the conventional correlation coefficient, and that the number of cases required to give a stable

estimate will be considerably increased. However, no estimate of the amount of difference is available.

The formulas which were most extensively used in the Aviation Psychology Program were formulas 2 on page 65, and 3 (or 3a) on page 66. We have already seen that the procedures described on page 68, and developed late in the war, should be substituted for the cases in which formula 2 is used. This has been done in some of the material presented in final summary reports. Some of the inadequacies of formula 3 to the situation encountered in aircrew selection have been discussed, but no better procedure has been discovered. Adequate treatment of this problem awaits further analysis.

SCORING FORMULAS

Scoring formulas for printed tests used in the AAF classification test battery were originally determined on a priori bases. In the case of relatively unspeeded tests, scoring formulas were assigned in accordance with the conventional procedure for correcting for guessing. The usual scoring formula was $R - \frac{W}{n-1}$, where n represented the number of answer choices for a given item. In the case of highly speeded tests, a substantially heavier penalty for errors was exacted. In a number of these tests the scoring formula R–3W was used. The heavy penalty for errors was used in order to place a considerable premium upon accuracy in those tests and as a practical procedure for giving comparable scores to individuals of comparable ability where in taking the test one individual placed more emphasis upon speed and the other upon accuracy.

Scoring formulas established as indicated above represented practical immediate operating procedures, but studies were also initiated to check upon these formulas empirically. For a number of tests the validities of the "rights" score and the "wrongs" score were determined separately. The correlation between these two scores was also determined. Using these values, it was then possible to determine empirically what weighting of "rights" and "wrongs" would give the maximum validity for a formula score.

In practice, it was found that in most cases test validity was relatively insensitive to changes in scoring formula over quite a wide range. This was due in some cases to the substantial negative correlation between "rights" and "wrongs." In other cases it reflected the small variability of the "wrongs" score. The insensitivity of test validity to changes in scoring formula had two practical implications. In the first place, it meant that scoring formula was not a highly critical consideration in test construc-

tion, so that a great deal of concern need not be given to it in the early stages of development of a test. In the second place, it meant that in order for empirical studies of scoring formulas to be of practical value, they needed to be carried out on extremely large groups of cases. Since the formula score validity varied only slightly over quite a wide range of change in the scoring formula, it could be expected that even small changes in the relative validity of the "rights" and "wrongs" score would produce drastic shifts in the optimum weight for the "wrongs" score in relation to "rights." As a result, personnel of the Aviation Psychology Program were inclined to depend, at least in part, upon rational considerations in assigning scoring formulas, as long as the validity of the rationally determined formula was not seriously less than that of the formula which had been empirically determined to be optimal.

In the discussion so far, attention has been centered on the efforts which were actually made to develop improved scoring formulas for tests used in the classification battery. It is appropriate to devote some time at this point to a more general consideration of the problem involved in getting maximum information from the successes and errors on a test. Interest in this problem was stimulated in the Aviation Psychology Program by the fact that a certain number of tests were discovered in which the "rights" and "wrongs" scores were essentially unrelated functions. This was true in particular of some of the highly speeded tests. For these tests it was found that a person who had a great many correct responses tended to have about as many errors as a person who had only a few correct responses. In other words, it appeared that speed and accuracy were somewhat independent functions and that the rapid individual might be either more or less accurate than the slow one. This permitted the obtaining of two separate scores from each test which were sufficiently independent statistically to permit of their being useful separate variables for use in research analyses.

In terms of immediate retest both "rights" and "wrongs" scores were often found to have reliability of the same general magnitude as formula scores. Whether this consistency in performance would be maintained over a period of time seems somewhat open to question. In a speeded test, speed and accuracy are to a certain extent conflicting goals, and performance in one direction can be improved at the expense of some loss in the other. The emphasis which is given to these two aspects of performance at a particular time of testing may be determined largely by temporary sets involving momentary interpretation of the test instructions. Over an extended period of time, an individual may show marked fluctuations in the emphasis he gives to each of these two goals. The

703315—47—6

score on either one of the single aspects may prove to be relatively much less stable, therefore, than a single score based upon an appropriately weighted combination of the "rights" and "wrongs." No evidence is available on the stability of separate "rights" and "wrongs" scores over a period of time.

Given that the "rights" and "wrongs" scores are sufficiently independent and sufficiently stable to make their separate analysis statistically meaningful, the question then becomes that of determining the most effective procedures for analysis. On theoretical grounds the most defensible procedure would seem to be to treat the "rights" and "wrongs" scores as two separate variables, each meriting analysis in its own right, and to include both of the variables in correlational studies. Each score would then be independently handled and independently weighted in determining regression weights. If the weights for the two scores were found to be different, then presumably the scores would be retained as separate variables in subsequent determination of weighted composite scores. This procedure has one fairly serious practical disadvantage in that it increases the number of variables to be dealt with in practical weighting operations. The compromise between maximum analytical value in a test battery and practical convenience which is involved here is entirely analogous to the one which is involved in the decision as to whether to include a number of related sub-tests in a single test score or to retain a separate score for each single sub-test. An intermediate manner of proceeding would be to keep the "rights" and "wrongs" scores separate during research analysis of the complete test battery, and thereby to determine the weighting of "rights" and "wrongs" which would make the test in question give its maximum contribution to the validity of the battery as a whole, and then to combine the two scores into a single one by means of the scoring formula which would give appropriate weights as determined by the previous analysis. This becomes practical only when the relative weights of the "rights" and "wrongs" scores are approximately the same for all job specialties for which the test is to be weighted. The separate analysis of "rights" and "wrongs" scores becomes particularly interesting and appropriate in exploratory studies for analytical investigation of the functions underlying test behavior, such as we find represented in the factor analysis approach to behavior.

In the use of "rights" and "wrongs" scores as separate measures in a test battery it becomes theoretically desirable to base each of the scores on a separate segment of testing. This is true because "rights" and "wrongs" scores based upon the same test period will ordinarily tend to have a negative correlation artificially introduced by the fact that each item which is correct

necessarily eliminates one possible wrong item. Scores based upon separate periods of testing can ordinarily be expected to be more nearly independent. From the practical point of view it becomes a question whether a specified amount of testing time could more advantageously be used broken into two separate shorter periods in order to achieve more independence of "rights" and "wrongs" or combined into a single period which would yield more reliable scores for both "rights" and "wrongs."

CHAPTER SIX

Obtaining Composite Aptitude Scores

The aircrew classification program was based upon the procedure of administering a number of varied tests to each subject and deriving from these an estimate of each man's aptitude for each of the various aircrew assignments for which he was a candidate. This approach lent itself very naturally to multiple regression techniques and those were in fact the procedures which were used. At this time it will be appropriate to describe the detailed procedures which were actually utilized in determining the manner of combining separate test scores into a single composite aptitude score and to consider alternative methods which might perhaps have been used.

PROCEDURES FOR DETERMINING BATTERIES, WEIGHTS, AND RECOMMENDATIONS FOR ASSIGNMENT

Attention will now be given to the procedures which were actually in effect in the program for combining tests, deriving aptitude scores from them and making recommendations for assignment. This involves, first, a consideration of the procedures which were used in arriving at a set of tests and weights for predicting success in a particular single category of aircrew training. It will then be appropriate to consider how the predictions for the several separate aircrew categories were combined into a recommendation as to the particular category in which a man should be trained. Thirdly, we shall give some attention to the bases for adding new tests to the test battery or deleting tests from the battery, and to the length of the test battery. Finally, it will be necessary to consider certain compromises which were made necessary by the practical demands for immediate testing for classification prior to the accumulation of adequate research data.

Prediction of Single Criteria

At this point we shall assume that data are available on the validities and intercorrelations of a battery of tests and that our immediate problem is that of deciding how to combine those tests

so as to give the best prediction of a particular criterion measure. We will not at the moment concern ourselves with where the battery of tests came from. In practice, after the aircrew program had once become established, there existed successive standard test batteries which had been used for classification and for which extensive validity data gradually became available. With these it was possible to combine, in various patterns, research tests for which some validity data were also at hand. The battery included in a particular analysis ordinarily consisted of some preceding standard classification battery as a nucleus, with the addition of one or more research tests. We shall not consider at this point problems as to the homogeneity of the data, particularly criterion data, for the different tests in the battery which we have specified above. In practice it frequently happened that validity data were not available upon the same sample for all the tests in a battery. Estimates of test validity were typically compounded from all available data on the validity of each of the tests in the battery under consideration. The estimates might be based on 30,000 cases for one test and less than 1,000 for another.

Given the battery as defined above, together with some estimate of validity and some estimate of the intercorrelations for each of the tests in the battery, determination of the weights to be given to the separate test scores in order to combine them into a single weighted composite score followed the general pattern of least squares determination of regression weights. At certain times during the program the standard Doolittle technique of computing regression weights was used. More frequently, however, regression weights were approximated by an iterative procedure. This procedure was approximately that developed by Kelley and Salisbury,[1] but was modified in certain details to take advantage of the additional computational efficiency which may be attained using a somewhat higher level of judgment than is required by the original Kelley-Salisbury procedure. The procedure is given in the Appendix.

In practice, exact regression weights were never used. A compromise was actually employed, made necessary by the use of the IBM test scoring machine with aggregate weighting board as an instrument for actually computing weighted composite scores. The aggregate weighting board puts limitations on the pattern of weights in two ways. In the first place, the blank which is used with the aggregate weighting board has space for a maximum of 30 tests in ten rows of three. All the tests in a single row receive the same weight. In the second place, the aggregate weighting board provides only for positive weighting of scores.

[1] Kelley, T. L. and Salisbury, F. S. An iteration method for determining multiple correlation constants, Jour. Amer. Stat. Assn., 21, 1926, pp. 282 ff.

The limited number of rows and spaces on the blank for the aggregate weighting board required some adjustments and compromises, especially when the battery of tests was being weighted for a number of aircrew specialties. In this case, a good deal of juggling of positions of tests on the blank was necessary in order to make it possible to approximate closely the desired weights for all the different composite scores. Since all the tests on a single row of the blank must receive the same weight, if they are weighted at all, there are marked limitations to the possible arrangements. It was only through masking out with masking tape certain board positions for certain score composites (and thus, in effect, weighting the test in that position zero for that job specialty) that a close approximation to the desired weights for the different jobs was possible.

The negative weighting of any test must be accomplished, using this equipment, by reversing the score scale for the test, that is, by subtracting all scores from a constant. Giving a test a positive weight for one aircrew specialty and a negative weight for another becomes very unwieldy in this case. It means that the test must be treated essentially as two tests and entered on the aggregate weight sheet twice, once with the original scoring scale and once with the reversed scoring scale. In general, analyses indicated that no large negative weights were called for by the existing pattern of validity coefficients and test correlations. The general procedure, then, was to use no negative weights but to compute the set of positive weights which, by the iterative procedure, reproduced as nearly as possible the validity coefficients of the component tests, and to base the actual classification upon this pattern of positive weights. The weights obtained in this way are equivalent to the regression weights which would result from a battery consisting only of the positively weighted tests. That is, the effect is essentially that of deleting from the battery any tests which would receive negative weights and basing the prediction upon the remaining tests in the battery.

In several instances, comparisons were made between the multiple correlation resulting from only positive weights and the multiple correlation which resulted when negative weights were also admitted. The gains from including negative weights were negligible in every instance. In the one instance in which a test was actually introduced with a negative weight, subsequent data indicated the negative weight to be of no value and it was withdrawn.

Elimination of negative weights had one other practical value. It simplified somewhat problems of public relations. It is quite difficult to explain, either to subjects or to the general interested public, why a man's rating for a particular job should be lowered

because he performed well on a test. The mathematics of the suppression variable is not easy to expound to the lay public.

Use of Aptitude Scores for Classification

Throughout practically all of the aircrew classification testing program, the final use made of aptitude test scores was as a basis for recommendation of classification for one or another of the three aircrew assignments, bombardier, navigator, pilot. In the later stages of the program, separate composite scores were obtained for fighter pilot and bomber pilot and for several types of gunnery training. Still later, scores were introduced for flight engineer and radar observer. However initial classification, as far as the aircrew classification program was concerned, continued to be in one of the first mentioned three categories throughout practically the whole war. A further enterprise was undertaken in the selection of gunners for assignment to B-29 aircraft, but this was, in effect, a separate enterprise applied either to those not qualified to receive training in any of the categories bombardier, navigator, pilot, or to groups already in training in enlisted specialties, and not a part of the single classification procedure for pilots, navigators, and bombardiers.

Composite aptitude scores were effective in assignment in two ways. In the first place, in order to be eligible for assignment to a particular type of training, a man was required to have at least a specified minimum aptitude score for that particular type of aircrew training. These aptitude scores were expressed in terms of standard score units on a scale from 1 to 9, in which 5 represented average and each scale unit covered a range of one-half standard deviation. This form of scale received the designation "stanine" at an early date in the program and the term became a part of the language of the program from that time.

The minimum qualifying stanines were determined by a number of practical considerations. In the early stages of the war, persons responsible for top policy were somewhat loath to disqualify anyone who had passed the preliminary screening with the AAF Qualifying Examination and the physical examination, and who had been accepted as an Aviation Cadet, from flight training in some one of the three specialties of bombardier, navigator or pilot. Therefore, initially no minimum score was set. Subsequently, data were accumulated which showed the effectiveness of the stanines for predicting success in training. At the same time, there developed a need for personnel to receive training as aerial gunners. Influenced by these factors, with undetermined weights, those in charge of policy acquiesced in a series of increases in the minimum qualifying scores. At the end of the war, a stanine of 7 or better was required to qualify for each one of the basic spe-

cialties of pilot, navigator, or bombardier. Many men disqualified from the above types of training were trained as aerial gunners.

In addition to providing minimum qualifying scores in terms of which certain men were disqualified from all types of aircrew training, the stanines provided a partial basis for determining the type of training for which a man should be recommended. The basis for recommendations included, in addition to stanine, strength of interest and willingness to waive first preference. In general, the procedure was to recommend men for the aircrew specialty for which they were qualified and for which they had the highest aptitude, excepting when this recommendation conflicted with the candidate's preference. In that case, preference was generally allowed to prevail unless the difference in aptitude scores was very pronounced or unless the candidate expressed himself as willing to be classified by his aptitude score rather than by his preference.

These procedures for determining for which of the several aircrew categories a man was to be recommended were clearly rule-of-thumb and had no mathematical basis. In this respect they contrast sharply with the procedures for determining weights in combining tests into a composite aptitude score. The weighting of tests was carried out with reference to an iterative approximation to the mathematically best combination of separate tests for predicting a training criterion, but the use then made of the weighted composite scores was based only upon practical considerations and professional judgment as to an appropriate way of combining the various different items of information. This contrast will be considered in somewhat more detail in a later section of this chapter.

Addition of Tests to the Battery

At this point we shall give some attention to the procedures which were used for determining when a research test should be added to the existing battery of classification tests. We shall assume for this discussion that a battery of classification tests had already been in use for some time and that the question which arose was whether or not a particular new research test should be added to the existing classification test battery. This question came up with regard to each new research test as soon as data were available with regard to its validity for any one of the aircrew specialties. In practice, these data were, in almost every case, data on validity for pilot training, because the usual administration of a research test to one or two thousand cases in a Classification Center yielded data for primary pilot training first and, in most cases, yielded sufficient data to be of significance only for the pilot category.

The data in terms of which judgments were made as to whether or not a research test should be added to the battery were the correlation of the test with the criterion of success in the particular aircrew specialty and the correlation of the test with the stanine for that specialty. These correlations had been corrected in most instances for curtailment due to selection of the individuals to receive training on the basis of stanine. Given these two correlations and the correlation of stanine with the criterion, it is possible to determine immediately how much the new test would add to the existing stanine if the two were combined with regression weights but without any internal changes in the existing stanine. When the tests already in the battery are correctly weighted, this gives a minimum estimate of the contribution of the new test to the multiple correlation when that test is combined with all the tests currently in the classification battery. It is a minimum estimate because internal changes in the existing stanine would be made only if they resulted in further increase in the multiple correlation and consequently could only result in still further increments. Of course, if the tests already in the battery are not optimally weighted, it is possible that the same increment in validity which is provided by the new test might be achieved completely or in part by re-weighting the tests already in the battery.

Partly for use in prevalidation analysis of research tests and partly to facilitate calculation of the amount that a new test would add to the validity of the stanine, a set of tables was prepared to show the validity required in a test if it were to add .01, .02, .03, .04, or .05 to the validity of another measure of known validity.[2] In this case, the other measure was the stanine and in preparing the table a value of .50 was used as an estimate of stanine validity. The table provided entries for different values of the test vs. stanine correlation. The values in the table were computed by the formula

$$r_{ck} = r_{ks}R_{c.t} \pm \sqrt{a\,(a + 2\,R_{c.t})\;(1 - r_{ks}^2)}$$

where

a = the specified increase in the multiple correlation
r_{ck} = the validity required of test k to achieve the increase a
$R_{c.t}$ = the multiple correlation of the battery, excluding test k, with the criterion
r_{ks} = the correlation of test k with stanine score when test k is excluded from the battery

It must be admitted that no satisfactory methods were available to determine the standard error of an increment in the multiple

[2] The formula and tables were developed by Lt. Col. A. Paul Horst.

correlation provided by the addition of a new test. In general, the statistics on the battery and those on the experimental tests did not include the same population, which fact made any estimate of the standard error of empirically determined increments still more difficult.

A new research test was considered worthy of more detailed statistical analysis insofar as the preliminary analysis showed it to make a substantial contribution to stanine validity, insofar as the sample for the experimental test was large, and insofar as the existing battery of tests for predicting success in the particular specialty in question was relatively poorly established or unsatisfactory.

These research tests which preliminary analysis indicated to be promising, in that they would increase prediction as single additions to the existing stanine, were typically added as additional variables to the matrix of battery intercorrelations and validities and a complete analysis was made of the tests in the battery and of the one or more promising research tests, using the iterative procedures for determining test weights described above. An outstanding advantage of this iterative procedure was that its speed made it practical to determine regression weights and resulting multiple correlation coefficients for a number of research tests as they were added singly and in combination to the existing battery. In such an analysis, the previous weights for the tests in the battery ordinarily provided a good initial approximation from which to make further iterations, and this procedure led to quite prompt convergence of the weights upon their final values. An examination of the multiple correlation coefficients resulting from the battery alone and the battery in combination with one or more research tests and of the regression weights for the several tests permitted a decision as to which, if any, of the research tests to add to the battery and which, if any, of the existing battery tests to drop upon the addition of research tests.

Although no exact mathematical standard was rigorously adhered to as the basis for adding a new test to a battery, during the last year or two of the war the working standard for selecting new tests for inclusion was that validity data based upon a minimum of 1,000 cases should indicate that the test would add .02 to the multiple correlation of the battery with the criterion. Each test which approached this standard was individually evaluated in terms of validity data and other statistical and practical considerations. As previously indicated, no applicable standard error formula for the increment in multiple correlation resulting from the addition of new tests to the battery was known; therefore it was impossible to determine how much regression an augmented multiple correlation could be expected to undergo in a new sample.

Length of the Aircrew Classification Battery

Throughout the war, the battery for aircrew classification was consistently maintained at about 20 tests. During most of that period, 6 of these tests were apparatus tests administered to subjects in groups of 4, each subject having his own copy of the apparatus to work on, while the remainder of the tests were printed tests administered to groups of 100 to 200 applicants at the same time. A complete day of each subject's time was required for the group testing session, while the apparatus tests required at least 2 hours on some other day. At this point it will be appropriate to consider the reasons for this extensive and rather elaborate testing program.

To begin with, it must be remembered that classification initially required estimates of aptitude for three distinct types of training —bombardier, navigator, and pilot—and that eventually other specialties such as flight engineer and radar observer were added. Though there was some overlapping of tests, the tests which were important for one specialty were ordinarily not the important ones for another specialty. Thus, the battery can in a sense be thought of as consisting of five or six tests for each aircrew specialty.

Some reduction in the number of tests might have been possible without serious losses in validity in any of the 3 aircrew positions, but it is difficult to get an entirely adequate evaluation of that point from statistical analyses. For particular samples of data, it was shown that predictions of all 3 aircrew positions could be obtained from one battery of 10 tests which gave correlations with the criteria differing from those obtainable from the complete battery of 20 or so tests by no more than .01 for any of the 3 aircrew categories. However it must be realized that these data were based upon the specific group for which the validity data were obtained. It must be anticipated that the values based upon only a fraction of the tests will show a marked shrinkage when applied to a new sample, a shrinkage which will be greater than when the regression weights are based upon all the tests. This point will bear a little elaboration.

If a large number of tests are given to a group of subjects and correlated with each other and with a criterion variable, the sampling fluctuations among the correlation coefficients will be sufficient so that it will practically always be possible to find some few tests which will give a substantial prediction of that criterion in that sample. However, in addition to any true relationship between the variables and the criterion, that prediction capitalizes upon the chance fluctuations in validities and intercorrelations of the variables. The smaller the number of variables

selected to be weighted, the more premium is placed upon chance favorable fluctuations in these particular variables retained and weighted. In general, the variables which are retained will be not only those that are most valid but also those which show the most favorable fluctuation from their true value in the particular sample. No accurate analytical formulation of this phenomenon is known, and it is not possible to estimate its magnitude with precision.

Combining Data from Various Sources

Ideally, in any program for the statistical analysis of test data and the determination of optimum weights for combining scores from a number of tests into a battery for the prediction of a criterion measure, data for all the tests being considered should be based upon the same large sample of cases. In this case, no question can arise as to the equivalence of samples available for different tests with regard to the population from which they were selected, the experiences to which they were subjected, or the criterion measures which were obtained upon them. Equality from test to test in these factors is guaranteed. In practice, however, with a real testing program in which testing time is limited and in which tests reach maturity over quite a period of time, this ideal cannot be achieved or perhaps even very closely approximated.

In practice in the Aviation Psychology Program, research tests were given for validation as they were completed and as time for experimental testing was available in the testing schedule. Validity analyses for each experimental test were carried out for the cases among those tested for whom adequate criterion data became available. For classification tests, validation was ordinarily carried out on much larger groups, often representing a complete class or several classes for a particular form of training. Intercorrelations were usually based upon Classification Center groups, which included not only those sent into several different types of aircrew training, but also those disqualified from aircrew training for low aptitude or other reasons. When statistical analyses of various research and classification tests were undertaken, it was necessary to assemble data from various sources. Ordinarily, test validities were estimated by making a weighted combination of all acceptable validity data on that test for the criterion being studied, and correlations were assembled from various sources. This procedure admitted of some heterogeneity of data from the different tests, but there seemed to be no alternative under the circumstances.

It was generally believed, though never convincingly demonstrated, that the above-mentioned heterogeneity was real and posi-

bly rather significant. If one research test happened to be validated upon a particularly favorable sample, it might be included in the test battery when such inclusion was not really merited, whereas another test which would truly have added to battery validity would have been rejected because it was validated upon an unfavorable sample. Stanine validity showed substantial fluctuation from one large group to another. It was rather generally felt that those variations were larger than could have been expected from sampling fluctuations alone, though no rigorous test of this point was ever made. Furthermore, the progressively higher standards for men sent into training as the war progressed made change in the elimination criterion very possible. In addition, there were, of course, the fluctuations in criteria and in validities arising from purely random sampling.

Some thought was given to the problem of making an adjustment to specific test validity coefficients based upon the validity of the stanine for the particular sample upon which that test was validated. That is, if the stanine validity was unusually low for the sample upon which a particular research test was validated, it seemed reasonable that the validity estimate for the test would be too low and that some allowance should be made for this. However, no satisfactory procedure was developed. Some purely intuitive allowance was made, in interpreting validity statistics, for stanine validity in the sample. In general, however, it may be stated that the problem of heterogeneity of data for different tests was one which was recognized but not solved.

Determination of Weights in Absence of Direct Empirical Data

The practical exigencies of military operations made it necessary in some instances to set up weights for use in classification before empirical data were available to make possible the type of analysis which we have discussed in the preceding sections. Research testing on an experimental basis was started in the AAF in a small way in the Fall of 1941 with the testing of 2,000 or 3,000 men on certain available tests. Shortly after war was declared it was decided that classification procedures would have to be put into operation immediately for differentiating between assignments to bombardier, navigator, and pilot training. Due to the length of time required for criterion data to mature and to the very short time lapse between the beginning of the research program and the requirement to start actually processing applicants, no data from the research program were available when the first battery of classification tests had to be established. Fragmentary data were available from testing in the RAF, RCAF, CAA and the Navy, and some information was available from job analyses in the AAF. In terms of these, it was necessary to establish a set

of weights for immediate use, subject to subsequent revision as empirical data became available. These initial weights were developed by psychological personnel in the light of testing in other organizations and of available job analyses. They were progressively revised as validation data became available from the direct results of the psychological testing program in the AAF.

Even after validation data began to become available from the AAF program, it still remained necessary to run somewhat ahead of existing data in various instances. For example, a test of pilot information was developed and introduced into one of the early test batteries. By the time validity data had been received for this particular test a number of other promising types of information test items had been suggested on the basis of validities of several different types of tests and had been incorporated in a revised pilot information test for use in the classification test battery. Collateral data were available for all the types of items in this revised test, in terms of their validities in other combinations and contexts. However, no empirical validity coefficient was available for the exact combination of items represented in the new test. It was necessary, therefore, to estimate the validity of this new test form somewhat impressionistically in terms of the collateral data. In almost all of the determinations of weights from analyses of a test battery, there were certain tests in which different types of estimation from data on similar tests or component test sections had to be used in arriving at the final validity value used in the correlational analyses.

Again, validity data were available most readily and sometimes exclusively for the more immediate criteria of success in early stages of training. However, observations became available from returned combat personnel and from visits of psychological personnel to combat theaters as to further abilities called for in the combat situation which did not show up in the more immediate training criteria. The degree to which weights based on statistical evidence of validity for training criteria should be modified by qualitative analyses and judgments of the distinctive requirements of operation at the combat level could be determined only on the basis of judgment by the personnel involved. In preparing some of the later batteries for the aircrew classification program, the actual regression weights against training criteria were tempered in some measure by such considerations. In the last analysis, then, it may be said that the set of weights used at any time for classification purposes involved some element of judgment as to the validity of one or more of the tests in the battery or as to considerations other than validity for training which should be taken into account.

Partial Criteria

As has been indicated, in practically all of the aircrew specialties for which personnel were being selected and classified in the Aviation Psychology Program, there were a number of available criteria. These consisted of criteria of different types and at different levels of training or operations. For example, in the case of pilot training, one criterion was supplied by elimination in training, and even this could be subdivided into the stages of primary, basic and advanced. In operational training a number of different criteria were available, including for fighter pilots air-to-air gunnery, air-to-ground gunnery, accidents, reclassification, and various types of ratings. Combat provided such criterion information as promotions, decorations, reported victories, casualties, and reclassifications. These types of criterion information all appeared to be in some degree relevant to judging the success of a particular individual, some more so and some less. The correlations among many of the separate criteria were found to be quite low. When several types of criterion scores are available, it is possible to determine the correlation of prediction tests with each of those criterion measures. One must then determine how to weight and combine the various partial criteria in determining the weights to be used in selecting personnel for assignment to this particular type of training.

No systematic procedure for carrying out this operation was reached in the Aviation Psychology Program. In general, the procedure was to base analyses on the most accessible criterion, namely that of success in training. As scraps of information were subsequently received on the more advanced stages of performance, some tempering of the weights based upon the training criterion was undertaken, but this was not upon any systematic or analytical basis.

It is possible to make an analytical, mathematical approach to the combination of partial criteria. The calculation of canonical correlations[3] provides a determination of the maximum prediction of a weighted group of criterion measures from a weighted group of prediction variables. The maximum prediction is achieved only when appropriate weights are assigned both to the predictors and to the criterion variables. Thus, when these weights are determined a mathematically unique solution is achieved for the maximum possible prediction of that group of criterion variables using that group of predictors. However, although this solution may be mathematically exact, there is some question as to whether it is practically meaningful. It seems doubtful whether

[3] Hotelling, H. Relations between two sets of variates, Biometrika, v. 28, parts 3 and 4, 1936, pp. 322-377.

there will be any particular correspondence between the weights assigned to part criteria in a mathematical solution and the judged importance of the part criteria in terms of the goals of training in that particular aircrew specialty.

The alternative to an analytical solution to the problem in terms of maximum prediction is a solution based on some composite of practical judgment. What seems to be required is some systematic way of assembling judgments of competent individuals as to the relative weight to be given to the various possible part criteria and of combining these judgments to yield a composite weighting scheme. Once the weights and intercorrelations of the partial criteria for which validation data are available have been determined, it becomes a relatively straightforward matter to determine the correlation of each test with that weighted composite and then to determine the appropriate regression weights for each test. The practical problems involved in this procedure appear to be those of picking an appropriate group of individuals to perform the evaluation and of developing a statement of instructions which will make the task maximally meaningful and clear-cut to them. The extensive use of judgment is inevitable when more than a single criterion variable is considered. The only question is how systematically the individual judgments shall be obtained and combined and to what extent the judgmental procedure shall be supplemented by mathematical analysis. Analytical procedures probably provide little direct support to judgment in determining how much to weight each partial criterion. Once that decision has been reached, however, analytical determination of the validities for the weighted composite criterion, and of the optimal weighting of test scores to predict this criterion, should provide a very relevant guide to the final decision as to test weights.

ALTERNATIVE METHODS OF DETERMINING QUALIFICATION AND ASSIGNMENT

At least two other approaches to the general problem of qualifying and classifying men for aircrew training may be considered. It will be appropriate at this point to take these up and indicate the considerations which led to their not being used as the procedure in the AAF aircrew classification program. The first of these will be designated the procedure of multiple cutoffs. In this procedure minimum qualifying scores are set for each of a number of tests separately, and those individuals are accepted as qualified who fall above the cutting score on all of the separate tests. A man is rejected who falls below the cutting score on any test. The second procedure may broadly be designated the clinical approach. This approach is distinguished by the fact

that the final judgment as to whether or not a particular individual is qualified for training or as to the type of training to which he shall be assigned is made individually for each man in terms of considerations which cannot be reduced to an objective mathematical formula. The clinical approach would ordinarily be expected to make use of data other than quantitative test scores, but it would also be possible to treat a set of quantitative test scores clinically and to base the decision upon whether the particular total pattern of scores obtained by an individual qualified him for a specific type of training.

Multiple Cutoff Procedure

This procedure may be examined both from the point of view of the logical assumptions which are implied by it and from the point of view of the practical operations which it would require if used in connection with an extensive battery of tests. It may be stated that the procedure was judged to be less acceptable than that of using a weighted composite score both from the logical and from the practical point of view.

Assumptions in Multiple Cutoff Procedure Compared with Those of Multiple Regression

The assumptions of a multiple cutoff procedure, in which a minimum score is established for each of the separate tests of a battery, may be compared with those of a multiple regression approach most simply in the case of two test variables. Let us assume that we have administered two tests to a population of subjects and that we wish to determine from the results on those two tests the procedure which will give us the most accurate prediction of success on some criterion, such as pass-fail in primary pilot training. The joint distribution of scores on the two tests can be shown on a two-way frequency scattergram. The joint frequency distribution will probably follow a pattern somewhat like that indicated in figure 6.1 shown below. In the case of multiple regression procedures we determine a linear combination of the two test scores such that a single aptitude score is computed. A minimum qualifying score established in terms of this aptitude score will be represented in figure 6.1 by line *a*. All individuals falling below and to the left of line *a* will be disqualified, and all those falling above and to the right will be qualified. The slope of line *a* is a function of the relative weight of the two tests in the combined aptitude score, and the position of line *a* is a function of the standard set to qualify for training. If separate minimum scores are established for each of the two tests, these will be represented in figure 6.1 by the lines *b* and *b'*. The effective difference in these two procedures is that those

703815—47—7

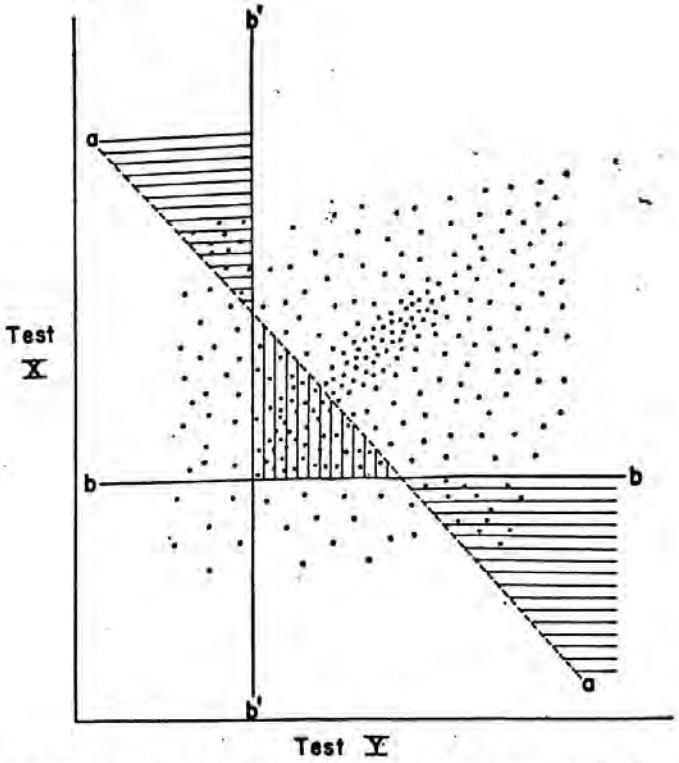


FIGURE 6.1.—Comparison of Multiple Regression and Multiple Cut-off Procedures for Personnel Selection.

individuals falling within the area indicated by horizontal lining are considered qualified in the first instance but not the second, whereas those in the area indicated by vertical lining are considered qualified in the second instance but not in the first.

The use of multiple cutoffs for different component tests would seem to be justified in those cases in which the relationship between test score and performance upon the criterion is conspicuously non-linear. If there is some point along the scale of performance upon a particular test below which all or most individuals fail in the job and above which few fail in the job, then a procedure which determines that point and establishes a sharp cutting score at that point undoubtedly has advantages. Insofar, however, as a continuous and approximately linear relationship exists between test score and success on the job, there is no basis for establishing a uniquely desirable cutting score on the particular test, and it is probable that the test can be used most effectively in a linear combination with the other test scores. In this connection it should be noted that one may expect some appearance of nonlinearity in empirical data from a limited number

of cases in many instances. The crucial issue is whether this same nonlinearity appears in subsequent samples. Any critical study of the results for multiple-cutoff procedures must be based upon cross validation of the procedures upon a new sample. Examination of the relationship of percent eliminated in training to test score for the various tests with which the AAF psychological program worked showed, for each test which had any substantial validity, a continuous relationship with progressively more eliminees at the successively lower score levels. In general, then, the data indicated that regression procedures were more appropriate than multiple cutoff procedures.

Practical Problems of Using Multiple Cutoffs

In the aircrew classification program, multiple cutoff procedures were rejected not only because empirical results indicating that the procedures were less appropriate to the data but also because of impracticality. Multiple cutoff procedures *may* represent a practical, and possibly an appropriate approach to the simple problem of selection when the selection is based upon no more than two or three tests. When, however, the number of tests increases to as many as 15 or 20, and when the problem is not merely one of establishing minimum qualifying scores but also one of accomplishing positive classification among the several aircrew specialties for which an individual may be qualified, procedures based upon successive cutoffs break down. The sheer burden of computational trial and error required to establish cutting scores for a battery consisting of a number of tests makes the procedure almost impossible from any practical point of view. If it were proposed to examine the appropriateness of no more than three different cutting scores on each of ten different tests, the total number of combinations for which the results would have to be analyzed would be approximately 59,000. A second major practical difficulty is that a series of cutting scores for a given aircrew specialty merely provides a basis for saying whether a person is or is not qualified for that particular specialty. It provides absolutely no basis for determining the one of several specialties for which he is best qualified. Since the original emphasis of the aircrew testing program was almost entirely upon classification and since qualification entered into the scheme only later as a subsidiary problem, it is clear that a procedure which merely provides a basis for deciding whether or not a person meets minimum qualifications for a particular assignment would have been entirely inappropriate.

Clinical Procedures

Throughout the course of the Aviation Psychology Program, pressure was repeatedly exerted to have use made of the clinical

approach in place of or in addition to the purely objective approach which had been adopted. It was urged that a rigid procedure of combining separate test scores, without any provision for evaluation or judgment of the pattern for a single individual, lost a good deal of valuable information which could have been used if skilled clinicians had been permitted to interpret the test results. Furthermore, it was contended that there were various types of data concerning the individuals which could not be reduced to objective test procedures but which could be obtained by the clinician through individual procedures of personal interview. In spite of these contentions, it did not seem feasible to use clinical procedures in the Army Air Forces classification testing program for reasons which are indicated below.

(a) An initial objection was that of requirements of time and training on the part of program personnel. Those espousing the clinical approach generally agree that the clinician needs time for a leisurely and unhurried evaluation of the individual and of the data pertaining to him. Furthermore, it is generally agreed that clinical procedures depend heavily upon the competence and experience of the individuals working with the subjects and rendering clinical judgments upon them. The aircrew testing program was, in the very nature of things, a mass program. There were periods in which certain of the testing units had to process as many as 500 candidates a day. The number of persons who, by even the most liberal interpretation, could have been considered qualified to make a clinical evaluation of the records of applicants for aircrew training would have been adequate to handle only a small fraction of this flow. The application of clinical procedures in any adequate fashion would have required a considerably larger total allotment of personnel that was available in the aircrew program and an enormously larger allotment of individuals with adequate clinical training.

(b) The clinical approach inevitably suffers from the fact that it cannot be standardized in any uniform way. This makes it no better than the great bulk of the individuals through whom it must be implemented. Even if it can be demonstrated that clinical procedures are valuable as applied by certain persons with specialized training and abilities, there is still no guarantee that those values can be maintained week in and week out by the group of relatively untrained individuals carrying on a routine procedure with numbers of cadets over a long period of time.

(c) Finally, it may be stated that the data which were available to the aircrew classification program provided no convincing demonstration of the practical effectiveness of clinical procedures for selecting individuals for aircrew training. A number of studies

were carried out on a relatively small scale, as tends to be necessary with studies of clinical procedures, in which an effort was made to validate certain types of ratings and subjective evaluations of personnel undergoing classification tests. These are described in more detail in Report No. 5 of this series of reports, but it may be stated in general at this point that the results were unpromising. There was little evidence to suggest that subjective evaluations of the type which could be made on the basis of Rorschach test, observation of performance while receiving psychomotor tests, informal observations during a rest period, and similar observations contributed to the validity of the battery based upon objective test scores. In certain instances, personnel officers or medical officers undertook to make exceptions to the then current minimum stanine to qualify for training. The exceptions presumably represented cases for which, in the clinical judgment of the officer in question, other factors compensated for the unfavorable stanine picture. There is no evidence that cases selected in this way graduated from training appreciably more often than would have been expected on the basis of their stanines.

PROBLEMS OF A UNIQUE CLASSIFICATION SYSTEM

It has been indicated in the introduction to this chapter that rigorous mathematical procedures were used only to determine the separate aptitude scores for each of the aircrew specialties. The procedure for determining to which specialty an individual could most advantageously be assigned was then a rule-of-thumb procedure which was based upon the difference between the single aptitude scores. The mathematical approach carried, therefore, only as far as setting up a set of selection devices for each of a number of separate aircrew specialties. That is, each stanine can be considered as a selective device for picking bombardiers, navigators, or pilots, as the case may be.

Classification also appears to present a group of problems which should be susceptible to approach in analytical terms. A pure problem of classification arises when we have N individuals to be allocated among N positions in k different categories, and it is desired to maximize the over-all effectiveness of the resulting organization. In the case which we are describing, there are no more men than jobs, so it is not possible to reject any individuals completely. In this case, it will be possible to make only limited use of the *absolute level* of the individual's aptitude for or ability in particular jobs; the critical factor will be *differences in level* of aptitude for or ability in the k possible assignments. The situation is further complicated, in the practical case, by the fact

that it will frequently be more important to approximate the maximum level of effectiveness in certain job categories than in others. Thus, in classifying AAF ground personnel it might have been thought more important to have the very best possible bombsight maintenance men than to have the very best possible cooks.

The classification problem has been somewhat generally formulated in the previous paragraph. We will try now to state it somewhat more explicitly, indicating the types of data which must be given to make an analytical statement of the problem possible.

Given:

(1) A limited number N of individuals available for job assignment.

(2) A limited number N of jobs which require to be filled, which are of k different kinds.

(3) A series of measures of individual aptitude or achievement.

(4) Data on the validity of each of the measures in (3) for each of the job categories in (2), together with the correlations among the measures and the correlations among the criteria.

(5) Weights to be attached to each job specialty, indicating the importance attributed to having maximum efficiency in that job.

Required:

A procedure for assigning the complete group of men in such a way that the weighted sum (by the weights in (5) above) of the aptitudes (or some function of them) of all the men in all the jobs shall be a maximum.

The above statement sets the classification problem in its pure form. The problem is clearly a complex one, and analytical approach to it will be very difficult even when satisfactory values can be established for all the "givens" listed above. However, this is the problem which must in fact be dealt with in many practical situations. It is currently dealt with in terms of professional judgment for which there is a minimum of systematic rational support. Any improvement in the analytical basis for differential assignment would seem to be a very worthwhile achievement. Although personnel of the Aviation Psychology Program were keenly aware of the limitations of current statistical procedures in providing the basis for a genuine classification program, it was possible to do little more than formulate the problems involved.

DESCRIPTIVE STATISTICS

In addition to the analytical statistical procedures which have been described in this and the preceding chapter, certain simple and effective techniques were needed to present the results of classification testing to a lay audience. A correlation coefficient has very little meaning to the reader unless he possesses a fair amount of statistical sophistication. Personnel in responsible positions who must make decisions with regard to the continuation or expansion of psychological activities are inclined to show a certain amount of impatience when they are faced with reports containing groups of correlation coefficients, regression weights and other unfamiliar statistical values. It becomes important, therefore, to devise effective ways of presenting results of testing research to statistically untrained personnel. From relatively early in the course of the development of the Aviation Psychology Program, psychologists devoted an appreciable fraction of their energies to the preparation of materials for distribution to nonspecialized personnel.

Most of the materials prepared for nontechnical use were graphic in nature. The most common of these graphic presentations was the bar chart. Since all composite aptitude scores were prepared on the 1 to 9 standard score scale (stanine score), they were in very convenient form for preparation of bar charts. The typical chart, of which a great many were prepared during the war, consisted of a series of bars, each showing the elimination rate at a particular stanine level. An example of one of these is presented in figure 6.2. This type of chart permitted a dramatic presentation of the different probabilities of success in a particular type of training for students with different levels of aptitude. Bar charts were prepared both for the composite aptitude score and for performance on the separate tests which went into the composite.

The type of bar chart which we have just described has certain limitations as far as providing practical information in terms of which administrative action can be taken. The question of the practical administrator is likely to take some such form as:

If we eliminate men below this specified level, what improvement may we expect in proportion completing training?

How many men will we have to recruit to fill quotas if these particular standards are used to qualify applicants? How many will we have to train?

Will this particular set of standards yield an appropriate percentage qualified for each of the different types of assignment?

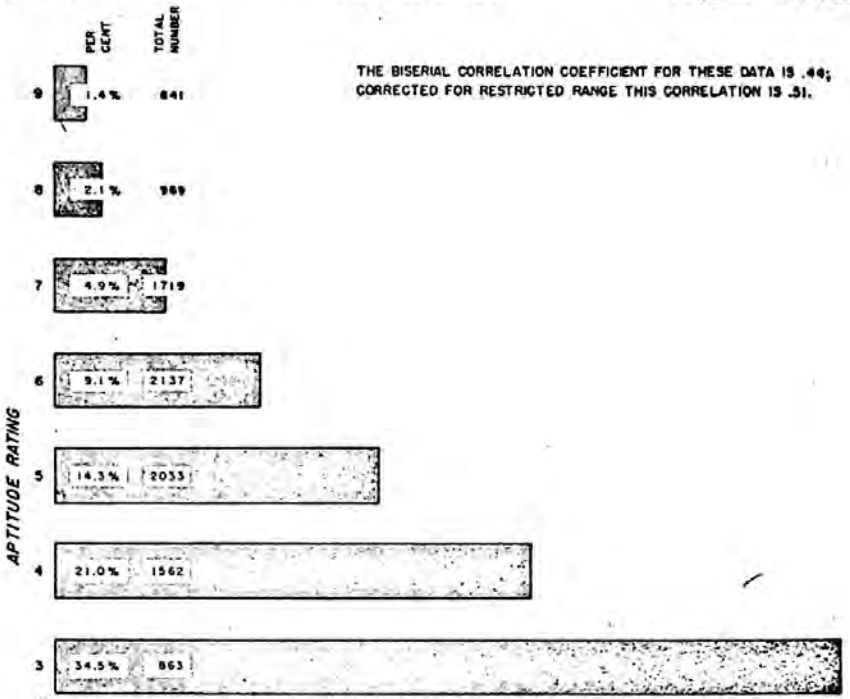
For practical planning purposes various tables were prepared, based upon the data on elimination rate at various stanine levels.

PILOT CLASS 44-E

Aptitude Ratings

PER CENT ELIMINATED FROM ELEMENTARY PILOT TRAINING FOR FLYING DEFICIENCY, FEAR, OR OWN REQUEST FOR EACH APTITUDE RATING
PILOT CLASS 44-E; ALL FLYING TRAINING COMMANDS COMBINED

10,161 CASES 1,243 ELIMINEES 12.2% ELIMINATED



THE BISERIAL CORRELATION COEFFICIENT FOR THESE DATA IS .44; CORRECTED FOR RESTRICTED RANGE THIS CORRELATION IS .51.

There Were 37 Cases With Aptitude Ratings Of 1 Or 2 Of Whom 43.2% Were Eliminated.

FIGURE 6.2

These showed the yield which could be expected out of a given group of men available for testing, in terms of the number who would be qualified to enter training and the number who could be expected to graduate if a given minimum score were required. Tables of this sort could be used in connection with the figures for flow of graduates which had been specified by higher headquarters so as to adjust flow into the different types of training in accordance with the required output and to yield maximum efficiency in terms of reduced training wastage.

CHAPTER SEVEN

Problems Associated With Reliability and Reliability Determination

NEED FOR DATA ON RELIABILITY

In general, the importance of information on the reliability of measuring instruments is thoroughly recognized. In fact, the significance of reliability has sometimes been overestimated, in that reliability in a test has sometimes been considered an end in and of itself rather than a necessary condition for obtaining significant relationships. However, it is worth spending some time upon a consideration of the particular values which data with regard to reliability have for a research program for the development of classification and selection procedures, together with a consideration of the difficulties which are encountered in defining and determining adequately the reliability of various measures. It is of some interest to see just what values are served by reliability statistics and to see in which contexts evidence on reliability is of only limited importance. We will find it appropriate to consider the significance of data on reliability, first, with regard to criteria, and secondly, with regard to analysis of test data.

Reliability Data in the Evaluation of Criteria

It goes without saying that one characteristic desired in a criterion is that it shall be reliable. The more reliable the criterion, the higher the correlation which may theoretically be obtained between that criterion and various predictive measures. However, it is not essential that the reliability of a criterion be high as long as the reliability is definitely greater than zero. Even when the reliability of a criterion is quite low, given that it is definitely greater than zero, it is still possible to obtain fairly substantial correlations between that criterion and reliable tests, and to carry out useful statistical analyses in connection with the prediction of the criterion. Since the range of values of the correlation coefficient falling between no prediction and maximum prediction is restricted for this relatively unreliable criterion, and

since the obtained correlations are consequently compressed into a narrower range, it will be necessary to increase the size of the population in order to obtain stable results. However, the unreliability of the criterion can be compensated for by such an increase in the size of the population.

Information about the reliability of a criterion is needed in the first place, to establish the fact that the reliability of the criterion is not zero. Unless this can be established at a reasonable level of confidence, further data based upon this criterion are likely to be quite ambiguous. In particular, negative results will be uninterpretable.

There were a number of cases in the aircrew classification program in which correlations were obtained between aptitude measures and existing criteria in training or combat and in which it was found that available tests gave no prediction of those criteria. In many of these cases, unfortunately, the nature of the criterion was such that no estimate of its reliability could be obtained. For instance, ratings were obtained on combat personnel from certain overseas Air Forces. In these ratings only a single observer evaluated each man, so that no estimate of consistency in making the ratings was possible. The available classification test data were found to have little or no correlation with ratings obtained in that way. However, general experience with ratings prepared by relatively untrained personnel and under conditions of minimum supervision leads one at least to entertain the possibility that the reliability of these ratings was essentially zero and that nothing could possibly have been found which would have correlated with them. In a case such as this, one is in the unsatisfactory situation of never knowing whether the failure to predict was due to the inadequacy of the test battery or the unreliability of the criterion.

A second value of reliability statistics for a criterion is to indicate what the maximum prediction is that could possibly be obtained for that criterion from a group of highly reliable tests. This last information is of significance in indicating what proportion of the predictable variance in the criterion has been accounted for by the testing procedures already developed and what portion remains to be accounted for by future tests still to be developed. It provides some guide as to the probability of significant gains from further research devoted to the prediction of the criterion in question and consequently some indication as to whether research can still profitably be pursued in that area.

Reliability Statistics in the Analysis of Test Data

Information with regard to reliability of a test may become of significance at two points in the sequence of test analysis and

evaluation. In the first place, data with regard to reliability are of interest during the initial stages of developing a new test. In the second place, reliability data are important for interpreting the correlations among a battery of tests.

When a new test is being developed we must be sure that the test achieves at least minimum standards of reliability. Other things being equal, the more reliable a test is, the more valid it will be. However, it must be remembered that the increase in validity of a test is not proportional to the increase in the reliability coefficient for that test but is rather a function of the square root of the reliability coefficient. Given a test with reliability of .64 and validity of .24, if the test is lengthened so that its reliability is increased to .81, we can only expect the validity to be raised to .27. Thus, we cannot expect large gains in validity by refinements to increase the reliability of an already moderately reliable test. In any event, in the initial development of a test the effort is to achieve as high reliability as possible without sacrificing other desirable test characteristics, so that the validity of the test may be attenuated as little as possible by chance error variance. In this connection, the gain from increased reliability must be weighed against the cost of increasing that reliability in terms of additional expenditure of testing time. In the case of a test battery the point of diminishing returns is ordinarily reached at a fairly early stage and beyond that point additional testing time devoted to a particular test will contribute less to the over-all validity of the test battery than will the same time devoted to some additional type of test materials. However, it would probably be difficult, if not impossible, to formulate a complete analytical statement of these relationships.

Our first use of reliability data, then, is as a guide in test construction. The data suggest whether details of testing procedure should be modified in the hope of obtaining a more reliable measure of the particular function being studied within the given period of testing time.

The second use of reliability data arises in connection with the analysis of the intercorrelations among a battery of tests. Our concern here is primarily with evaluating the uniqueness of each test as an independent contributor to the test battery. Within a given test battery, we can think of the variance of a single test as being divisible into three fractions. One fraction is variance which is common to that test and to other tests in the test battery. This variance of a single test is predictable from the other tests in the battery and its amount is given by the square of the multiple correlation between the test in question and all the rest of the tests in the battery. A second fraction of variance in a given test score is error variance; that is, variance which

is specific to that administration of that test and which could not be predicted even by administration of another form of the same test. The amount of variance of this type is indicated by the reliability coefficient for the particular test. The third fraction of variance for a test, and the fraction in which we are most particularly interested, is the fraction which is genuine, systematic variance in individual behavior (is predictable from day to day and from one form to another of the particular test) but is variance which cannot be predicted from the rest of the tests in the battery. This fraction of variance represents the individual and unique contribution of the test in question to the total battery. Insofar as a particular test score can be predicted from the other tests in the battery, it contributes nothing new of its own and can increase the total predictive power of the battery only through increasing the reliability of the composite score. Only insofar as the test has unique variance can it extend the proportion of variance in the criterion which is covered by the battery as a whole. In the evaluation of a particular test, therefore, knowledge as to this uniqueness is of crucial importance. This third fraction of variance, systematic variance unique to the particular test, is defined as the difference between the two fractions which we have previously discussed. In order to determine it adequately, knowledge of the reliability of the test is indispensable. Reliability is important, therefore, in providing one statistic which must be used to evaluate the uniqueness of each of the tests in a test battery, and consequently the possibility that the test in question may extend the range of human behavior covered by the battery.

A third, and perhaps minor, significance of data on test reliability is the role they play in the determination of the maximum correlation between a test or battery of tests and a criterion. This maximum is, of course, a function of the reliability of the test or battery on the one hand and the reliability of the criterion on the other. In practice, the reliability of the test or battery of tests is likely to be enough higher than the reliability of the criterion so that test reliability becomes a minor factor in determining the possible correlation.

FORMULATION OF THE CONCEPT OF RELIABILITY

It is a mistake which has sometimes been made by students in the field of tests and measurements to conceive of reliability as a single, universally defined concept which has the same meaning at all times and places and to all individuals. The reliability of any measure is always defined by a set of operations, and there have been various different sets of operations used to define reliability which gives substantially different results. It must be

recognized that variance in performance upon a test arises from a great many different sources. A particular set of operations for defining reliability treats certain of these sources of variance as sources of error, and others as sources of true variation in performance upon the measure in question. The different operational definitions of reliability disagree somewhat as to the sources of variation which they put into the one or the other category. It will be necessary therefore, to examine the sources of variation in human behavior as it shows itself upon a particular measure, and to determine which of these are logically to be considered sources of error variance and which sources of systematic variation in behavior.

Sources of Variance in Test Scores

Source of variance in test scores can be broken up into a number of different categories. Table 7.1 on the following page gives such an analysis of variance in test performance. This analysis is probably not complete with regard to all the minor categories, but it does indicate the major categories of variance and some of the specific elements which may occur within each. The question becomes one of deciding which of the types of variance are to be considered systematic variation among the individuals in performance upon the test in question and which are to be considered sources of error.

With regard to certain of the categories of variance, there will be no disagreement either from the logical point of view or from the point of view of the different sets of operations which define reliability. These categories will be allocated in the same way by all the different operational definitions of reliability. For example, all of the variance under category I will certainly be treated as systematic variance. The pertinent characteristics of the individual with regard to general traits and abilities affecting this particular test are certainly a source of systematic variance in test performance. Though it may be desirable so to design our test that variance attributable to I-A and I-B is reduced to a minimum, since ordinarily we are not interested in obtaining a measure of the individual's "test wiseness" or ability to read, these factors, as well as the general traits underlying test performance, *do* produce variance which is a systematic feature of the test.

The variance under II-A would also uniformly be considered systematic variance in test performance, although there might be some question as to how narrowly the test should be defined and therefore as to how much of II-A 2 should be considered as general.

There will be similar agreement that the variance described under categories IV-B and V represents error variance and is a

source of unreliability on the particular test in question. Any definition of reliability will include as error variance the pure "chance" variations in performance which show up in moment to moment fluctuations in efficiency of performance and in the sheer hazard of guessing answers to particular questions.

The sources of variance listed under II-B, III, and IV-A are less clearly attributable to either the category of systematic variance or the category of error variance. We shall see that different operational definitions of reliability treat these sources of variance in different ways and that there are logical considerations which make sometimes one and sometimes the other treatment more reasonable.

TABLE 7.1.—*Analysis of Possible Sources of Variance in Performance on a Particular Test.*

I. *Lasting and general characteristics of the individual.*
   - A. General skills and techniques of taking tests.
   - B. General ability to comprehend instructions.
   - C. Level of ability on one or more general traits, which operate in a number of tests.

II. *Lasting but specific characteristics of the individual.*
   - A. *Specific to the test as a whole (and to parallel forms of it).*
     1. Individual level of ability on traits required in this test but not in others.
     2. Knowledges and skills specific to particular form of test items.
   - B. *Specific to particular test items.*
     1. The "chance" element determining whether the individual does or does not know a particular fact. (Sampling variance in a finite number of items.)

III. *Temporary but general characteristics of the individual.* (Factors affecting performance on many or all tests at a particular time)
   - A. Health
   - B. Fatigue
   - C. Motivation
   - D. Emotional strain
   - E. General test-wiseness (partly lasting)
   - F. External conditions of heat, light, ventilation, etc.

IV. *Temporary and specific characteristics of the individual.*
   - A. *Specific to a test as a whole.*
     1. Comprehension of the specific test task (insofar as this is distinct from I B)

2. Specific tricks or techniques of dealing with the particular test materials (insofar as distinct from II A 2)
3. Level of practice on the specific skills involved (especially in psychomotor tests)
4. Momentary "set" for a particular test

B. *Specific to particular test items.*
1. Fluctuations and idiosyncrasies of human memory.
2. Unpredictable fluctuations in attention or accuracy, superimposed upon the general level of performance characteristic of the individual.

V. *Variance not otherwise accounted for (chance):*
"Luck" in the selection of answers by "guessing."

The variance under II-B represents the variance due to the particular sample of items which we have chosen to represent the total area being measured by the test. Any test which is made up of discreet items of knowledge or skill chosen to represent the large and almost unlimited set of possible tasks within an area introduces this problem of sampling. In general, the correspondence between knowledge of one item in a field and knowledge of a different item will be less than perfect, so that tests made up of different sets of items will correlate less than perfectly because of the particular sample of items of which each is composed. The only operation for determining reliability which does not recognize this sampling of items as a source of error is the determination of reliability by retesting with identically the same materials. This operation becomes acceptable, then, only insofar as one has no concern about variation in the sampling of items from the area to be studied. This is reasonably the case when the materials with which one is concerned are very homogeneous as, for example, in the case of a series of simple perceptual judgments. When the material is sufficiently varied so that a problem arises as to the representativeness of the particular sample of items, operations for determining reliability should be such that the variance due to the sampling of items is permitted to be classified as error variance.

Under III are listed various factors causing variation in individual performance from day to day or possibly even from hour to hour. These represent changes in the individual, whether from general conditions of health and emotional adjustment or from more specific learnings resulting from particular training which he may be undergoing, either in or out of the test situation. It must be recognized that a particular test of an individual is a test of that individual as he is at a particular time. For some measures the variation in the individual from one time to another

may be an important source of variance in test performance. Insofar as these fluctuations are general, they will affect a variety of different performances in the same way. Any operations for the determination of reliability which make use of some device for splitting up score on a particular test given at a particular time ignore these sources of variance as a factor producing differences in individual scores. In fact, reliabilities based upon a single period of testing tend to allocate daily variations in the individual to the systematic variance rather than to error variance, and these fluctuations then serve to raise rather than lower the estimate of reliability.

If our purpose is to determine how accurately a test administered to an individual at one time can predict his performance at some other time, operations based upon a test at a single time are likely to give a spuriously high estimate of this type of reliability. The only acceptable procedure for determining that reliability which is defined as the ability to predict the individual's performance at some future time is to retest the individual after a lapse of time. If, as indicated above, the sampling of items is also a significant source of error variance, this retest should presumably be with a parallel form of the original test. A parallel form should be defined for this purpose as one which conforms to the same specifications as the original test in terms of content, difficulty level, and standards of internal consistency of items, but which, within those specifications, selects a random sample of items.

For some purposes we may not be interested in knowing how accurately we can predict an individual's performance at a future date. In particular, when our interest is to analyze the correlations among a group of tests, all given at the same time, the appropriate definition of the reliability of each test is in terms of the individual as he existed at the time, because the correlations among the tests are based upon the individual as he existed at a particular time. In this case variance from day to day is not involved in the intercorrelations and need therefore not be taken into account in the estimate of reliability. In this instance, reliability may appropriately be defined in terms of some procedure for dividing the test into parts which are then correlated with each other.

Some further issues are involved, particularly in connection with the types of variance which were listed under IV-A. There appear to be a number of relatively temporary factors influencing only performance on a single test. These are factors which we consider to be specific to that performance and to a limited period of time. They involve such things as "getting the hang of" instructions, developing an efficient technique for taking a test and

the like. These are in considerable measure hypothetical rather than demonstrated factors. Their importance has not been shown empirically and is not as immediately apparent as that of most of the other factors which we have considered.

It was believed that if factors of temporary set and of grasp of instructions were important they would show up particularly in highly speeded tests, in tests where the task was quite novel to the subjects, and in tests which involved rather complex situations and instructions. One effort was made to study this matter by comparing reliabilities from two separately timed parts of a test (a) when the two parts followed one another immediately and (b) when several hours filled with the administration of other tests intervened.[1] Four tests were selected which were believed to show to a rather high degree the characteristics described above. They were given to groups in counterbalanced order, each group receiving one test without interval and one with. The reliabilities without interval were .58, .40, .65 and .85; with interval they were .58, .41, .60 and .82. On the average, the reliabilities after the interval were lower by about .02. Though these results are consistent with the existence of some variance in performance due to the types of temporary sets and procedures which we have suggested, they do not indicate that variance to be great in amount.

When variance of type IV-A is significantly present, immediate retests with a particular type of test will yield inflated reliability coefficients because of temporary factors influencing only that test performance. This variance will always be unique to that test and will give it the appearance of having some genuine specific quality which, in truth, has no lasting or permanent significance.

Evaluation of Operations for Reliability Determination

There are a number of different sets of operations which have been suggested or used at one time or another for computing an estimate of reliability. It will be appropriate at this time to consider each of these in relation to the sources of variance which we have just discussed, in order to see how the various types of variance are disposed of in the different sets of operations. The different approaches to reliability determination may be classified as follows.

a. *Retest*
   (1) With same test form
      (a) Immediately
      (b) After an interval

[1] See Report No. 5, Printed Tests, Ch. 3, for a fuller report of this study.

(2) With an equivalent test form
  (a) Immediately
  (b) After an interval

b. *Sub-divided Test*
  (1) Alternate items
  (2) Alternate groups of items
  (3) First and second half
  (4) Equivalent halves

c. *Analysis of Variance Techniques*
  (1) Hoyt
  (2) Kuder-Richardson

These will now be considered in turn in the light of the operation which they employ and the logical consequences of these operations.

Immediate Retest with Same Test Form

Evaluation of reliability by immediate readministration of a specific test form and the correlation of the two resulting sets of scores will in effect exclude from the estimate of error variance and include in the estimate of systematic variance the types of variance listed in I, II, III, and IV A in table 7.1. We may question the logic of including as systematic variance the variance in categories II B, III, and IV A. In some cases, the variance in category III may reasonably be accepted as systematic variance. This is the case when we are interested in evaluating correlations among a group of tests administered upon the same day. Since day-to-day variations in this case represent a systematic factor producing covariance among the tests, they may reasonably be considered a source of systematic variance when estimating the reliability of the separate tests.

There are some types of tests for which the variance in categories II B and IV A may be insignificant. The variance in category II B will disappear as the test items become very homogeneous, as, for example, in a series of psychophysical judgments or in a test with many items of simple arithmetical computation. Variance associated with the particular sample of test items will also disappear in simple repetitive motor tasks involving reaction speed, coordination, and the like. An aspect of variance of type II B is memory of the specific test items and of the previous response to them. This is an objection to immediate retest for any test in which single items are sufficiently distinctive in character so that memory of them is a probable occurrence. Variance of type IV A will probably be unimportant for all familiar types of test materials and for any tasks which do not require maximum speed and attention.

It would seem, in summary, that an immediate retest with the same test form is a satisfactory operation for estimating reliability only (a) when day to day fluctuations in performance are a consideration of no importance, i. e., when the reliability data are going to be used in conjunction with other data gathered at the same time and (b) when separate items are homogeneous and not individually identifiable.

Retest After an Interval with Same Test Form

This differs from the procedure just described in that variance under factor III and IV A is allocated to error rather than systematic variance. At the same time the factor of memory of the responses to specific items is minimized. Therefore, when day-to-day variance is to be considered as error variance (i. e., when it is desired to estimate the consistency of performance from one time to another) and when the materials are sufficiently homogeneous so that the selection of specific test items is not a significant source of variance, this procedure seems quite appropriate. Types of homogeneous test materials, as indicated above, include simple motor tasks of speed, coordination and the like, series of psychophysical judgments, and very simple and numerous mental tasks. Examples of the latter would be tests of cancellation, substitution, simple numerical operations and the like. Only empirical evidence can demonstrate which types of materials are so homogeneous that two repetitions of the same test will show no higher correlation than two equivalent test forms. In the absence of this evidence, it is always safer, and always at least as satisfactory on logical grounds, to use equivalent forms. The problem of length of interval between testing, which is discussed in a following section concerning equivalent forms, is also relevant here.

Immediate Retest with an Equivalent Form

This procedure differs from immediate retest with the same form in that variance due to category II B is now correctly allocated as error variance and, since the items are not the same in both tests, there is no problem of memory for specific items. The issues which arise concern variance in categories III and IV A. If these are significant sources of variance, it may be desirable to adopt a procedure which allocates the variance to error, rather than to systematic variance. We have already indicated that for use with intercorrelations based upon the same day of testing variance in category III may reasonably be thought of as systematic variance. If, as is suggested by the experimental comparison of correlations resulting from immediate and delayed retesting referred to previously, factor IV A is not important

as a factor in most printed tests, the operations involved in immediate retesting with an equivalent test form constitute an acceptable definition of test reliability.

Delayed Retest with an Equivalent Form

This set of operations differs from the set just discussed only in the delay, which has the effect of allocating variance in categories III and IV A to error variance. In this case, the only variance which is treated as systematic variance is that in categories I and II A. This provides a rigorous definition of reliability in terms of the accuracy with which the test will predict performance on other measures of the same function at some other time. This definition seems appropriate whenever we are interested in an index of accuracy of the test as a measure of a particular type of function over a period of time.

The problems which arise in connection with specifying more precisely the operations in this definition concern the time interval between tests and the definition of "equivalent" forms. Upon investigation, it may be found that the correlation between forms is a function of the time interval between the two test administrations, dropping somewhat as the interval increases. If this is the case, the interval between test administrations should be specified in any report of reliability data, and an interval should be selected which is in some meaningful way related to the practical purposes for which the testing is to be used. That is, if the test is to be used to predict success in flying training six months after testing it would be appropriate to report retest reliability with an interval of six months between test and retest.

The definition of an "equivalent" form also presents certain problems. How specifically and exactly must one form of a test duplicate another in order to be considered equivalent? This would seem to be a question of defining what it is the test is supposed to measure. Presumably, the function of a test is defined in terms of a set of specifications for the construction of the test. In the case of an educational achievement test, for example, the specifications might outline the content in terms of the number of items to be allocated to certain broad areas, the important points within each area about which questions might be asked, the distribution of item difficulties, and the manner of selecting items with regard to internal consistency. Within the limits set by the specifications, the particular test form should be a random selection of items of knowledge and skill. Two tests may be considered equivalent forms if they both conform to the same set of specifications. Forms should not be required to be parallel item by item and should correspond only in that they each represent a random selection of items within the same limits

of content area, difficulty level, item format, etc., which are set up in the test specifications.

Sub-divided Test

Evaluation of reliability by sub-dividing the items on a test, after it has been administered as a unitary whole, is comparable to an immediate retest with an equivalent form, except that the two scores are not based upon separately-timed performances and that the items entering into the two scores may be in some degree interspersed. The criticisms which apply to the immediate retest apply here, together with certain additional ones which arise from the lack of separate timing and the mixing of items.

A split-test reliability gives a completely meaningless index of reliability in any test which depends primarily upon speed. In this case, score on any group of items depends primarily upon their position within the test, i. e., upon whether the individual had an opportunity to attempt them. A reliability based upon odd vs. even numbered items is spuriously high because opportunity is systematically equated between the two part scores. A reliability based upon first vs. second half of the test is meaningless in that the individual only has opportunity to score on the second half of the test insofar as he finished the first. If score on a test is partly, but not entirely, a function of speed the spurious and distorting effects which have just been described continue to operate, but to a lesser degree.

A second possible spurious source of reliability lies in moment-to-moment fluctuation in performance. If quality of performance fluctuates during the taking of a test, this will mean that error of measurement will tend to be correlated for successive items. The procedure of systematically assigning alternate items to the two halves of the test will tend to equate the effect of these fluctuations on the two half scores, so that they will operate to inflate rather than reduce reliability coefficients.

The computation of reliability on the basis of odd vs. even numbered items introduces the possibility of one other type of spurious effect tending to inflate reliability coefficients. If successive items in a test tend to be more alike than a pair of items taken at random from the test (or from that particular section of the test) the reliability coefficient will be inflated. This will happen because of the systematic assigning of alternate items to the two half scores. The type of error which has been described may arise in tests in which several items are based upon the same material, as in reading comprehension tests where several items are based on the same passage or interpretation of data tests where several questions refer to the same map or table. In any case such as this, items should probably be subdivided

by larger units, so that the part scores are based upon odd vs. even numbered passages in a reading test, or the like.

The question of how to split the items in a test into two part scores is analogous to that of how to build up two equivalent forms of the test. Presumably each half test should conform to and be representative of the specifications in terms of which the total test was made up. That is, there should be equal representation of the major types of items which were specified in planning the total test. Within that outline, items or groups of items should be assigned at random.

Analysis of Variance Techniques

All the procedures which have been described so far require that two separate scores be obtained, and that the correlation between these scores be determined. They have differed in the manner of defining the two scores. Where the length of the tests which are correlated is different from that of the total test which is finally to be used, the standard Spearman-Brown correction formula is used. This formula is

$$r_{AA'} = \frac{nr_{aa'}}{1 + (n-1)\, r_{aa'}}$$

where $r_{aa'}$ = reliability of test of length a

and $r_{AA'}$ = reliability of test of length A = na

The procedures which are now to be considered approach reliability directly through the analysis of the variance in test scores. These procedures apply to the analysis of a single test and were developed as a replacement for those procedures which require the subdivision of the test into two separately scored parts. The analysis of variance approach provides a unique value for the test reliability which is not dependent upon the particular sampling of items which is included in each test score. It thus avoids a certain ambiguity of definition which arises whenever one particular way of subdividing a test into two subtests must be selected from among all the possible ways.

Analysis of variance procedures are based upon a comparison of error variance within a test score and total variance of a group of subjects taking the test. Estimates of error variance can be obtained from the inter-item correlations, item-test correlations, or from simpler statistical values which are equivalent to these under certain limiting assumptions,[2] or from a subtraction of examinee variance and item variance from the total

[2] Kuder, G. F. and Richardson, M. W. The theory of estimation of test reliability. Psychometrika, 2, 1937, pp. 151-160.

variance.[3] No effort will be made at this point to present either the detailed procedures or the derivation for the above approach. The reader is referred to the cited references for that material. It will be appropriate, however, to consider the assumptions which are made in this method and the way in which the different fractions of variance which were outlined in table 7.1 are allocated.

The analysis of variance approaches make one assumption which is implicit also in the split-test procedures. This is that each individual has an opportunity to attempt each item, i. e., that speed is not a factor. It is only when the item has been attempted by all subjects that item difficulty, item intercorrelations, and the like take on meaning. These procedures were conceived in connection with the purely "power" type of test and it is only in this case that they are applicable.

The, further assumption is made that the non-error variance in each item covers the same factors in human behavior as that in every other item, i. e., that the test is completely homogeneous so that any subdivision of items into two parts is as reasonable as any other. In mathematical terms, this means that the rank of the matrix of item intercorrelations is unity. This procedure would not be applicable, therefore, to composite tests made up of more than one type of content.

In certain of the simplified procedures which have been set up for facilitating computation, additional assumptions are made. One of these has been equality of item difficulty. These additional assumptions are recognized as producing an underestimation of the reliability obtained from the more laborious computations, but some evidence has been offered to show that the difference is not large.

Referring to Chart I, analysis of variance procedures will allocate to systematic variance that variance in categories I, II A, III and IV A. The same questions with regard to categories III and IV A may be raised here as were raised in connection with the immediate retest with an equivalent form. These are, of course, in addition to the points which have been made in the immediately preceding paragraphs.

SPECIFIC PROBLEMS IN RELIABILITY DETERMINATION

So far the discussion has been a general one presenting logical considerations involved in estimates of reliability. It will now be appropriate to turn our attention to a number of specific problem situations which arose in the work of aviation psychologists and for which some set of operations for estimating reliability had to be established.

[3] Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 6, 1941, pp. 153-160.

Reliability of Speeded Tests

A very large number of tests involve to some degree the factor of speed. At one extreme are tests upon which every individual could perform perfectly if given sufficient time, so that the only source of differentiation between individuals is in the speed of their performance. From this extreme, tests range to the other at which the conditions of testing allow ample time and any variation among individuals is purely in terms of their level of performance upon the task in question. Many tests fall somewhere in between the two. Insofar as the element of speed is important for a test, it is impossible to obtain an adequate estimate of reliability from a single test administration with a single time limit. The odd-even procedure for determining reliability gives an inflated estimate because there will necessarily be, for any individual, approximately the same number of not-attempted items among both the odd and the even items. The first half vs. second half procedure underestimates reliability because all of the variation in number of items attempted tends to appear in the second half of the test. The various Kuder-Richardson formulas are not appropriate because they provide no basis for taking account of items which were not attempted. The only legitimate procedure is to administer the test in two separately timed parts. Whenever a research test is constructed in which it is suspected that the element of speed may be important, it should be constructed in two equivalent parts which may be separately timed, so that an adequate estimate of reliability may be obtained.

Reliability of Psychomotor Test Involving Progressive Learning

In the case of most psychomotor tests, the problem of estimating reliability is complicated by the fact that individuals show a steady improvement in performance on the test from beginning to end. If the improvement were uniform for all individuals, no particular problem would be involved because the various learning curves would be approximately parallel. An individual's standing relative to his group at one point in the practice curve would be approximately equivalent to his standing at other points. Insofar, however, as individuals show widely different rates of improvement, a problem is introduced. The problem is one of defining what we mean by reliability in connection with a learning task. Do we mean the accuracy with which a person's position has been determined at a particular point in his learning curve, or do we mean the accuracy with which a score at one point in the curve characterizes his relative performance at some future time? If the latter, how large a span of the remaining curve do we undertake to include in our prediction? Five minutes, five hours, or five months?

In practice, we are usually interested either in obtaining a prediction of performance after a considerable period of time without intervening practice on the test or in evaluating a test score obtained on a particular date from the point of view of its relationships to other tests. For the former purpose, presumably our best estimate of reliability is an actual test-retest reliability with a substantial time interval between the two testings. For the latter, we must choose between various possible methods of subdividing the initial test period. The choice is usually between breaking the test period up into a number of small fractions and computing the correlation between the odd and even numbered fractions or dividing the test up into two halves and computing the correlation between the first and second half. The chief objection to odd-even reliability in this case is that the assumption of independence of errors of measurement in successive fractions may not be justified. Successive fractions or trials are likely to be subject to the same chance influences, so that chance fluctuations affect both the odd and even scores in the same way and serve to increase spuriously rather than decrease the obtained estimate of reliability. The objection which may be raised to the first half vs. second half estimate of reliability is that learning factors may have influenced the second half score differentially for different individuals so that the consistency of individual performance is somewhat concealed by differences in individual rate of learning. However, if the total is taken as the unit, score on the test is in part a function of differences in rate of learning and it would seem that these are legitimately a source of unreliability in our estimate of individual performance. It is believed therefore that the more legitimate estimate of reliability for a psychomotor test is that in which score on the first half of the testing period is correlated with score on the second half. This point of view differs from that which was effective during the war, so that most estimates of psychomotor test reliability are based on odd vs. even trials. It may be anticipated that these will be biased in the direction of being too high.

Reliability of Tests with an Element of Discovery

In standard test statistics it is assumed that a test is made up of relatively homogeneous elements and that these continue to present the same task to the individual as he takes the test. With familiar and more or less standard types of test items this is probably essentially true. When the task which is presented the subject is relatively novel, however, we must expect him to show a certain amount of learning with regard to techniques for solving the problems which the test presents. Changes in technique may appear as gradual increments of skill; they may appear as

relatively sudden "insights" or "hypotheses." The individual may discover new clues or a new focus of attention, or he may more or less suddenly "get the idea" of what the test is about.

When the element of discovery and sudden change in the level of performance on a test become important, the usual techniques of reliability determination become to a large extent meaningless. An individual's score depends upon the point in the testing procedure at which he "caught on." A reliability based on odd vs. even items or trials may be spuriously high, just as in the case of a speed test, because the gains accruing from this "catching on" are evenly split between the two halves. A first vs. second part reliability will be lowered if insights came at different times during testing for different individuals. If insight was an all-or-none matter at the time instructions for the task were presented, the reliability may even in this case be a function of the presence or absence of insight rather than of skill which would be shown on the task after the basic idea of the test had been comprehended.

It seems possible that both some of the apparatus and some of the printed tests involved this insight type of factor. In one psychomotor test, improvement from very poor performance to almost perfect performance took place quite suddenly for particular individuals, and the individual's score appeared to be determined more by the point at which this improvement took place than by a general level of performance. In some of the experimental paper and pencil tests it seems probable that the most critical part of the test was the instruction period, and that the performance of the individual reflected at least in part the degree to which he "got the idea" from the instructions rather than the particular proficiency which he had in the skills required for actually performing the test. It is difficult to present any objective or convincing evidence of the operation of the type of factor which we have been discussing. Its occurrence as a possibility can best be evaluated by examining the instructions for certain of the more involved types of tests.

No really satisfactory technique is known for measuring the reliability of a test where score depends primarily upon presence or absence of insight into the test task. The various types of split-test reliability tend to be unsatisfactory for the reasons which have been discussed. Obviously, a retest is not a satisfactory solution, because once insight has been obtained it is no longer possible to present the individual with the same task which he faced when he first approached the test. In some cases it may be possible to produce an equivalent task calling for a new but comparable insight. In general, however, there will be no guarantee as to the comparability of the new task and the old. A low

relationship between the two may not mean unreliability of performance, but rather inadequacy on the part of the test constructor in producing an equivalent task. The problem is one for which no completely satisfactory solution appears to be available.

Reliability When Result of Performance Is Known

In the typical paper and pencil aptitude or achievement test the individual responds to a large number of successive test items and then turns in his paper or answer sheet. If reliability is to be estimated from a retest, the retest is ordinarily administered before the individual has any opportunity to observe or profit from his performance on the first test. The individual is provided with a minimum of information as to the adequacy of his reactions, so that he has little or no opportunity to improve his later performance by a study of his earlier errors. A number of situations were encountered in the Aviation Psychology Program in which this was not the case. Particularly in performance measures of proficiency in aircrew duties, it was often possible for the individual to observe his errors in the initial attempt at the task and to modify his subsequent behavior in the light of the observed error. For example, if performance in landing a plane was being used as an indication of pilot proficiency and the individual was making repeated attempts to land the plane, his errors on the initial attempt to land were likely to be painfully obvious to him. If his initial landing was high, so that he stalled out and came in with a terrific bounce, he could not help but know the nature of his error and he would naturally tend to try to avoid that same error in his next landing attempt. Or again, where the log of a standardized navigation mission was being used both as a measure of navigational proficiency and as a part of the training program for improving the skill of student navigators, the errors which the student made upon a particular flight normally served as the basis for criticism and remedial instruction, so that there was a systematic program for teaching the student to avoid repeating these same errors.

Insofar as either spontaneous individual observation or planned instruction brings the subject's attention to his specific errors between one testing and the next, a systematic factor is introduced tending to reduce consistency of performance. In an extreme case we might say that the individual who commits an error upon one occasion can almost be counted upon not to commit that error the next time. It is entirely possible that in carefully avoiding his previous error he may commit some other and that his overall performance viewed as a total may still be consistently either good or bad, but even this is not necessarily the case. In the extreme then, the effect that we have just discussed would pro-

duce negative correlations between successive measures and in a less extreme case reliability would be systematically reduced.

In the case of spontaneous observation of error by the individual, the effect of this observation could probably be reduced by increasing the interval between successive tests. That is, the individual would be expected to remember more accurately an error which he had committed on an immediately preceding landing attempt and to correct it upon another attempt five minutes later than he would from one day to the next. In cases where spontaneous observation of behavior is important, therefore, it is probably desirable to let an appreciable interval elapse between successive testings which are being used to determine reliability. In those cases in which systematic instruction is being given, however, lapse of time does not seem to provide a solution. This is the case because that time will be filled with instruction centered around the previous errors, and the changes produced in the individual by this instruction may be expected to be more fundamental and lasting than those resulting from his own incidental observation. It may be necessary to sacrifice some instructional values for research purposes if a really adequate estimate of reliability is to be obtained.

Independence as a Factor in Reliability of Ratings and Subjective Evaluations

As tends to be true in all cases in which ratings are used as a measure of individual proficiency, independence of the ratings presented a critical problem in the Aviation Psychology Program as far as reliability determination was concerned. Really adequate estimates of the reliability of rating procedures were difficult to obtain. Lack of independence in the rating by different raters arose from two somewhat different sources. On the one hand there was a problem of actual collaboration among raters. In the practical administrative task of preparing evaluations of personnel, collaboration is not necessarily bad, since it is at least possible that a joint evaluation prepared cooperatively by two or three men working together is as accurate as the average of the two or three separate sets of ratings. When it comes to evaluation of the reliability of ratings, however, any collaboration of this type is fatal to obtaining a true estimate.

Where ratings were not closely supervised and were something of a burden to administrative personnel, it is entirely possible that officers prepared them in a somewhat perfunctory manner and talked them over with each other as they did their work even if instructed not to. A difficult and more serious problem is one not of direct collaboration but rather of indirect contamination through what we may speak of as the man's local reputation.

Where a student is taught by different instructors or flown by different check-pilots, it is probably typical that these men informally talk over the student with one another and that a somewhat generalized picture of the man develops at that particular station. This reputation may carry a good deal of weight, so that later evaluations of the man are only in part a function of his actual performance and are in considerable part a function of his general reputation as that is known to the rater.

A good deal of evidence suggests the importance of the second factor which we have just discussed. For example one analysis indicated that the reliability of the composite of check-ride grades given in primary flying training was approximately .80. The correlation between grades in primary and grades in basic flying training has been shown to be only in the .20's. A similar relationship was found to exist when correlations of different ratings in a single phase of fighter-pilot operational training were compared with correlations between ratings in the two phases.

Special administrative precautions can eliminate direct collaboration between raters evaluating individual performance in a given school. The more subtle effect of general reputation, however, cannot be taken care of in this way. There is probably little or nothing which can be done about this within a given school. It presents a general, recurring problem in the evaluation of ratings as a measure of proficiency and suggests that probably the reliability of ratings must be evaluated in terms of consistency between successive levels of training.

Within Versus Between Missions Reliability

In a number of types of aircrew criterion data it was possible to obtain split-test reliabilities either by splitting the separate performance within each mission into two parts or by splitting the successive missions into odd vs. even missions. Data have been reported for each of these two procedures and in general the procedures yielded strikingly different results.

In any given test performance we recognize score as resulting in part from variance in the basic ability of the individual and in part from variance of different types extraneous to the basic ability of the individual in question. Many of the sources of extraneous variance are more or less uniform within a given flight mission. These are such factors as pilot, plane, instruments (bomb sight, astrocompass, etc.), weather, and temporary condition of the subject. In the case of the procedure which splits the alternate gunnery attacks, bomb drops, or other units of behavior within a given mission and allocates them in part to each of the two half scores, all of this variance due to external conditions is equated between the two part scores and made to work

in the direction of producing an appearance of reliability. The reliability which results in this way is of course spurious. If factors of the type we have just mentioned are important sources of variance in test score, the spurious effect may be quite substantial. We must conclude, therefore, that the procedure of subdividing performance on a single mission is indefensible and that the results from such a procedure are essentially meaningless.

When reliability is determined by correlating odd vs. even missions, the variance of the type which we have just described is ordinarily randomized and tends to reduce rather than inflate the obtained reliability. This is, of course, appropriate since variance of the type which we have discussed is error rather than systematic variance. The procedure of determining reliability by correlating performance on one mission with performance on another is ordinarily satisfactory. One situation must, however, be recognized in which this last procedure gives a systematic underestimate of the actual reliability which exists. In those cases in which factors such as pilot, plane, and the like are systematically shifted from one mission to the next, rather than allowed to vary at random, a tendency has been introduced to produce negative correlation between successive missions. That is, the individual who had the best pilot on one mission will necessarily have a poorer one on the next and vice versa. In certain experimental designs systematic rotation of background factors was resorted to in order more nearly to equate the conditions for each student during the total experimental period. In some of these cases reliability estimates comparing one mission with another are systematically too low.

CHAPTER EIGHT

Certain Problems In Correlational Analysis

SIGNIFICANCE OF INTERCORRELATION IN PREDICTION PROBLEMS

Research officers in the Aviation Psychology Program entered upon their work with a lively awareness of the practical importance of test intercorrelations, and this awareness was confirmed and strengthened by results obtained within the research program. Test intercorrelations were not only a matter of continuous practical concern, but also the basis for a good deal of theoretical discussion. In this chapter a few of the practical and theoretical considerations are elaborated.

The practical importance of test intercorrelation can be illustrated quite simply and dramatically. Let us assume that we have several tests, each of which has a correlation of .30 with a criterion. Let us next assume that all the intercorrelations of these tests are first .00, then .10, then .30, and finally .60. The values of the multiple correlation which can be obtained from various numbers of tests which meet these specifications are shown in table 8.1. Examination of this table makes it abundantly clear that the test intercorrelations are a factor of prime impor-

TABLE 8.1.—*Effect of Intercorrelation on Multiple Correlation.*

Multiple correlation resulting from different number of tests, when validity of each test is .30 and intercorrelations are uniform and at several different levels.

No. of tests	*Size of intercorrelations*			
	.00	*.10*	*.30*	*.60*
1	.30	.30	.30	.30
2	.42	.40	.37	.34
4	.60	.53	.44	.36
9	.90	.67	.48	.37
20	(*)	.79	.52	.38

*It is mathematically impossible for 20 tests all to correlate .30 with some measure, and still have zero intercorrelations.

tance in determining how good a prediction can be obtained from a battery of tests. In general, the contribution which any single test can make to the effectiveness of a battery for predicting some criterion is a function on the one hand of the correlation of

the test with the criterion and on the other of the correlations of the test with other tests. Those tests will have high positive weights in the regression equation which have high validities and low intercorrelations. Tests with low validities and high intercorrelations may sometimes be valuable as suppression tests, that is, as negatively weighted tests. A suppression test is one which overlaps the non-valid variance of some valid test, so that when the suppression test is negatively weighted it serves to partial out the non-valid variance, and thus purify the measure of the valid factor. However, almost no clearcut examples of useful suppression tests were found in the Aviation Psychology Program.

When the problem under consideration is that of weighting an existing battery of tests so as to predict a single job criterion, standard procedures for computing multiple regression weights (as discussed in Chapter 6) will take appropriate account of test intercorrelations, and no special further thought need to be given to the problem. When, however, the problem is one of planning a program of test development research, one of selecting a battery of tests to provide differential prediction among a number of job specialties, or one of streamlining a test battery so as to obtain maximum predictive efficiency for one or several job specialties within a limited amount of testing time, the problems of intercorrelation are more complex, and less susceptible to direct analytical solution. These problems will be discussed further in the following sections.

MAJOR TYPES OF TESTING PROJECTS

Obviously, the usefulness of any test in a testing program is a function of its validity for the criterion or criteria which the research worker is trying to predict. However, there is room for a good deal of variation in the emphasis which is given to the simple factor of validity. On the one hand, it is possible for a test constructor to make this the one central consideration in his test development activities. On the other, considerations of validity in single tests may be to a substantial degree subordinated to considerations of correlation with other tests. The desirable balance between these two emphases will depend upon the use to which the test is to be put.

The uses of tests for the evaluation of personnel with view to job assignment fall into three general patterns. Tests may be used

a. As a screening device to qualify personnel for assignment to a single job or type of training. *(Selection)*
b. As a multiple screening device to qualify personnel for assignment to some one or more of a number of jobs or types of training. *(Multiple selection)*

c. As a device to determine to *which one* of a number of available jobs or types of training a person should be assigned. *(Classification)*

The second and third categories above are not exclusive, and a test battery may be used simultaneously both to qualify and to classify. During most of the war this was the situation which prevailed in the Aviation Psychology Program. With emphasis shifting from one to the other function as time went on, tests were used both to disqualify those who failed to meet the minimum standards for any of the types of aircrew training and to determine the type of training which should be recommended for each individual who qualified for more than one.

The evaluation of test development procedures must be made in terms of the purpose for which the tests are being developed. The amount of emphasis on obtaining maximum validity as opposed to obtaining minimum intercorrelations, and on the development of complex job analogy tests as opposed to simple tests of human functions depends in some measure upon the one of the three categories presented above with which we happen to be concerned. We shall now examine the qualities desired of a test battery in the light of each one of the three types of use and see what implications this has for test development.

THE USE OF A TEST BATTERY FOR SELECTION

When a battery of tests is being developed to qualify personnel for assignment to a single job, the one quality which is desired (in addition to purely practical ones such as economy, convenience, etc., which are somewhat outside the scope of the present theoretical discussion) is validity of the battery in terms of an adequate criterion of performance on that job. The only thing that matters is the correlation of the final composite score derived from the battery with the criterion. Test development activities are focused on making this correlation a maximum.

With this in view, development of tests which resemble the criterion task both in content and complexity is a natural approach. One can readily see the rationale for having the content of the testing situation resemble as nearly as possible the actual duties on the job. One reasons, probably soundly, that the more nearly the test approaches the job or some phase of the job, the more predictive test performance will be of job performance. Thus, for pilots one constructs motor coordination tests which use an airplane-type stick and rudder, for navigators one constructs a table-reading test using data on drift, airspeed, and the like. One plans the test so that it may measure a certain type of general function, but so that it also measures it with the specific materials and in the specific type of situation which is likely to occur in the job

in question. In this way, one hopes that factors of specific content as well as factors of general function may give the test validity for the job being studied.

A related but somewhat different aspect of construction of tests for a single job is the tendency to make the tests complex. A job will ordinarily be complex, requiring the individual to do a number of different things, often at the same time. In the effort to reproduce these conditions as nearly as possible, the test is likely also to become complex. In the Aviation Psychology Program this was seen in the large number of complex coordination and pursuit tests which were adopted or developed for pilot selection, requiring the individual to use simultaneously a number of controls and to respond to a number of signals and cues.

The defense which is made for complex tests based on the materials of the job is that individually they tend to have high validity for the job for which they were particularly tailored. This seems often to be true. It is further urged that the use of material related to the specific task on the one hand and the introduction of complexity of function on the other introduces validity which could not be covered by *any number* of simpler and more general tests of mental functions. This may also be true, though it is harder to demonstrate conclusively. It is certainly true that there were among the tests developed for aircrew selection a number of complex tests which had validity beyond that which could have been achieved by combinations of the simpler tests which were available at that time. This does not, of course, exclude the possibility that a number of simple tests *could* ultimately be found which would collectively account for all of the validity of such a complex test as the Complex Coordination Test, or the like. However, it seems safe to say that within the scope of test development of the Aviation Psychology Program, the available simple tests would not collectively have given as valid a prediction of single job criteria in pilot and probably navigator training as was obtained from the battery using complex tests.

In the case of a pure selection battery, the only criticism that can be leveled at the complex type of test, developed with an eye only to validity and without regard to intercorrelations, is that each job-analogy type of test will tend to have relatively high correlations with other tests built on the same basis, and that consequently relatively little gain can be obtained by adding to an existing battery other tests of this same type. However, no ready basis appears available for answering the crucial question of whether the final multiple correlation resulting from a well-planned battery of complex tests will be higher or lower than that resulting from an extensive battery of simpler tests. One effort was made in the Aviation Psychology Program to compare

the results from experimental batteries of these two types as applied to a new job specialty, but limitations in the adequacy of the test batteries and meagerness of the data prevented the study from being at all conclusive.

THE USE OF A TEST BATTERY FOR MULTIPLE SELECTION

When tests are being used for purposes of multiple screening, the theoretical situation is not essentially different from the above. In a sense one has two, three or more testing batteries all given at the same time, each of which is used to predict success in a particular job category. Each battery could, in theory, be developed in complete independence of the others and carried to the point of giving the best possible prediction of a single job criterion. In practice, however, the need for efficient use of limited testing time precludes the development of such parallel independent batteries. They might be possible if the number of job specialties were only two or three, but they would become hopelessly inefficient, unwieldy and time-consuming with a larger number of jobs. In that case it becomes necessary that each test be used in the prediction of success in several jobs.

As it becomes necessary to use a single test in the prediction of success on not one but several jobs, it obviously becomes less defensible to design the test in terms of the duties of a particular job. Of course, the test may still be conceived as functioning *primarily* for a *single* job, being used incidentally in the prediction of success in other job specialties insofar as it is found empirically to predict those job specialties.

Thus, in the battery of aircrew tests certain tests were conceived of primarily as pilot tests, others as navigator tests, and still others as bombardier tests. However, each test was weighted for any aircrew specialty for which analysis indicated that it should receive weight. Furthermore, in expanding the use of the battery to additional specialties, the tests developed for pilot, navigator, and bombardier provided the basic battery for the new specialty. Thus, the tests which had been developed for selection of bombardiers, navigators, or pilots were reweighted when it became necessary to select flight engineers. This was, however, considered something of an expedient pending the development and validation of tests more specifically directed at predicting flight engineer criteria.

If tests are to be designed less in terms of the activities of a particular job, they must be designed more in terms of general categories of human behavior. The approach to test development in terms of aspects of human behavior starts off with the search for and definition of behavior categories. Categories may be drawn to a large extent ready-made from the language of the in-

troductory psychology textbook or of everyday speech. In this way one may set out to build tests of "judgment," "attention," "observation," "memory," and the like. However, the verbal labels provide only starting points for test construction, and as the necessary set of operations is undertaken to translate the categories into usable tests, certain difficulties are likely to arise. In particular, one is likely to find that tests which purport to measure the same category of behavior, and which should be functionally nearly the same, have only a moderate relationship, and that tests which purport to measure different categories, and which might therefore be expected to be independent, are in fact related to a fairly substantial degree. In other words, test scores often do not organize themselves into sharply defined clusters corresponding to a priori categories.

This had led to the effort to refine categories in terms of the empirical results of test intercorrelations. The effort at developing refined and more useful categories depends in every case upon obtaining the matrix of test intercorrelations. There is great diversity, however, in what is done with the correlations by different workers after they have been obtained. On the one hand, the correlations may serve primarily as material for sophisticated inspection, in terms of which the tests are re-evaluated and hypotheses are formulated as to new test operations which are expected to provide more nearly unique and uncorrelated tests. These tests will then serve to define separate and distinct dimensions of human behavior. On the other hand, the same goal is sought through the complex series of operations involved in factor analysis. Factor analysis undertakes to resolve the test correlations into a number of independent components, and through rotations of these components to identify each with both some nameable aspect of behavior and some test or tests.

The goal of the refinement of categories is to get a set of categories which are mutually independent and collectively inclusive. Insofar as this goal can be achieved, the resulting set of categories, and the set of measures to represent them, will have both logical and practical advantages. On the other hand, it will be simpler to think and talk about a set of categories all of which are separate and distinct, rather than in varying degrees interrelated. From the practical point of view, independence of the several tests will contribute to the efficiency of the battery to be used for multiple selection. Each test will measure a new aspect of human behavior, with a minimum of duplication of what has been covered in other tests. A maximum scope of human behavior will be evaluated within a given period of testing time. Insofar as predictions must be made for a number of jobs involving a variety of types of duties, and insofar as it is consequently necessary to

evaluate many different aspects of behavior, this non-overlapping in different tests may be a matter of great practical importance. It becomes a matter of theoretical importance in connection with the problem of classification which we shall consider next.

THE USE OF A TEST BATTERY FOR CLASSIFICATION

In the case of multiple selection, which we have just been considering, the goal remains the relatively simple one of attaining maximum accuracy in the prediction of each of the job specialties taken singly. Since it is not possible to design a separate battery for each job specialty, compromise with the ideal must be made, and some loss in accuracy of prediction of single job categories is tolerated in order that the prediction of others may be improved. The practical goal is that the average prediction of all the job specialties, with appropriate weight given to the importance of each job, be a maximum within the limits of time and facilities which are available for testing.

As soon as the task becomes one of classification, an entirely new element is introduced into the goal of testing. In a strictly *classification* program, it is assumed that each man *must be used* in some one of the available specialties, and that the purpose of testing is to determine his relative fitness for each of the different duties. At this point, we are no longer interested primarily in *level of aptitude* for single jobs, since we must use even the poorest men somewhere. We are now interested in *differences in level* between different jobs. It is no longer sufficient to predict success in job A accurately and to predict success in job B accurately; we must predict difference in success between jobs A and B accurately. This means that we must be interested not only in the validity of our test or composite score for job A and the validity for job B, but in the degree to which the predictions are differential. This can be clarified by reference to the familiar formula for the correlation of sums and differences.

If we let

A = score predicting success in job A
α = actual success in job A
B = score predicting success in job B
β = actual success in job B

then (A — B) will be the predicted difference in success on the two jobs and (α — B) will be the actual difference in success. The validity of the differential prediction will be the correlation between these two differences and will be given by the formula

$$r_{(A-B)\ (\alpha-\beta)} = \frac{(r_{A\alpha} + r_{B\beta}) - (r_{A\beta} + r_{B\alpha})}{2\sqrt{1-r_{AB}}\ \sqrt{1-r_{\alpha\beta}}}$$

703815—47—10

From this formula it can be seen that the critical factor in differential prediction is that the difference in the validity for a particular job specialty of the score used to predict *that* job specialty and the various scores used to predict *other* job specialties be a maximum. The actual level of the complete set of validity coefficients is not important. It is the amount of difference among them that matters. It is of interest further that for a given set of validity coefficients, high correlation among the prediction scores makes for more rather than less validity of differential prediction. Obviously, low correlations between prediction scores will tend to go with great differentiation among validities. However, it is of interest to note that there is no virtue in trying to reduce the intercorrelations of prediction scores artificially. In particular, for valid classification purposes it is desirable that the errors of measurement for the different predictions be *as highly correlated as possible.* Error of measurements is thereby held constant for the different job categories, and cannot operate to produce invalid discrimination between them.

Returning to procedures for test construction, it has been shown that for a *classification* program the measure of the success of a test battery lies in the differential validity of the several predictions for the several jobs. Two composite scores can have different validities only insofar as they measure different functions. The only validity of the battery for classification purposes, therefore, lies in the difference in function measured by the different scores. A test is of value only if it permits the differentiation of some function from other functions.

Let us suppose we have tests such that each represents a pure measure of some trait of behavior, the traits being isolated and refined by the best available statistical techniques and professional insight so that they are as nearly unrelated and as psychologically meaningful as may be. In this case, we may expect the validity of a particular test to differ sharply from one job criterion to another, since the trait will be important for some job assignments and not for others. This differentiation will not be blurred by any other functions entering into the single test scores. Tests of this sort will permit a maximum differentiation of the validity of each composite score for the different job specialties, since it will be unnecessary to include in the composite score for one job any of the secondary sources of test variance which would tend to give the composite relatively more validity for other job specialties than the one which it is designed to predict.

In the case of highly complex tests, the reverse will tend to be the case. The complex test is likely, by its very complexity, to include elements which have validity for a number of aircrew

specialties. A combination of several such tests may be quite valid for the specific job for which the composite was assembled, but it is also likely to be quite valid for other jobs as well. Thus, its value as an instrument of *differential* assignment is reduced. In general, then, it appears that independence of separate test scores is of particular importance for the task of classification.

No mathematical solution is known for the problem of classification as it has been outlined above. Consequently it is not possible to state the conditions for maximum effectiveness in classification with any exactness. By the same token, it is not possible to indicate, except as has been done in a very general way, the techniques and points of emphasis which will be most fruitful in producing an effective classification battery. In general, it seems that the emphasis will need to be much more on pure tests of distinct functions and much less on valid tests for specific aircrew duties.

CHAPTER NINE

Sources and Control of Error Variance in Test Scores

INTRODUCTION

The performance of any individual on a test or group of tests is in part a function of more or less general knowledge and ability on the part of the individual being tested, in part a function of the specific sampling of tasks which he is called upon to do, and in part a function of a variety of incidental environmental factors beyond the control of the individual. These environmental variations are in part individual, unique, unidentifiable, and unmeasurable. Here we refer to such individual, incidents as having a cold, having slept poorly the night before, having received an upsetting letter from home, having been distracted while the instructions were being presented and the like. Such factors are highly individual and represent sources of inaccuracy in measurement about which the testing organization can do very little. In part, the environmental variations are recurring, identifiable and possibly measurable circumstances which differ between individuals or between groups. This category would include variations in temperature and humidity, variations in details of procedure from one testing unit to another, variations between specific examiners, and variations between copies of the same apparatus test. Since these last factors can be identified and individuals can be segregated into subgroups within which a particular factor was held constant, it is possible to make statistical studies of the effect of these factors, and, where they are of practical importance, to develop experimental or statistical procedures to allow for them.

Fluctuating environmental influences are of importance in testing because they lower the accuracy of measurement, and consequently the validity of resulting scores. It should be noted that the lower accuracy will only influence that type of reliability index which is defined by retest at another time when the specific set of environmental conditions no longer holds. If split-test or immediate retest procedures are used, the environmental conditions will ordinarily be the same for both part scores, and the

variance in conditions will appear as systematic rather than as error variance.

The ideal procedure for dealing with environmental variations in testing is obviously to eliminate them by control of the testing situation. It is to this end that all the precautions to achieve uniform testing are introduced. In the Aviation Psychology Program, every effort was made to maintain uniformity of conditions from man to man, day to day, and unit to unit. Procedures for standardization developed progressively during the war. Originally, each of the three Psychological Research Units worked up its own standard routine of procedures and detailed operations. More and more centralized direction of testing operation was gradually obtained, and eventually a complete and detailed Standing Operating Procedure was set up by Headquarters, AAF Training Command. The Standing Operating Procedure specified in detail procedures for administering, proctoring and scoring tests, and tallying test results. Verbatim instructions were supplied for administrators of both group and individual tests. The individual test instructions specified not only what the test administrator was to say but what he was to do and how much demonstration he should provide the subject. In addition, uniform procedures were specified for apparatus calibration and apparatus maintenance, and for auditing and checking all scoring and conversion procedures.

Even when all possible precautions are observed to reduce the effect of environmental variation, it is still possible that significant effects may remain. It is desirable, therefore, that periodic checks be carried out to determine to what extent significant effects exist, with a view to controlling them or correcting for them. The remainder of this chapter discusses some of the types of factors which were analyzed in the Aviation Psychology Program. In interpreting these results, it must be remembered that these analyses are of the variation which remained in spite of all efforts to maintain uniform and controlled conditions. That some factor was not a source of significant variance in the Aviation Psychology Program is no indication that it would not have been if conditions had been less well controlled.

VARIATION BETWEEN TESTING UNITS

Some general clue to uniformity of testing conditions or lack of it could be obtained by comparing mean scores for different testing units for each test, month by month. The interpretation of these data was somewhat ambiguous, since it was not possible to guarantee that applicants arriving for testing at the different units were equivalent in aptitude. Different units tended to serve rather different geographical areas, and there may have

been other systematic differences between the populations tested. Repeated experience led personnel in the Aviation Psychology Program to have a healthy skepticism as to the applicability of the assumptions of random sampling to the populations with which they had to deal. However, the plotting of monthly means by units appeared to throw some light upon uniformity of testing conditions from unit to unit. This was possible partly through analysis of the internal relationships of test scores. That is, a unit which yielded consistently high scores on one test but

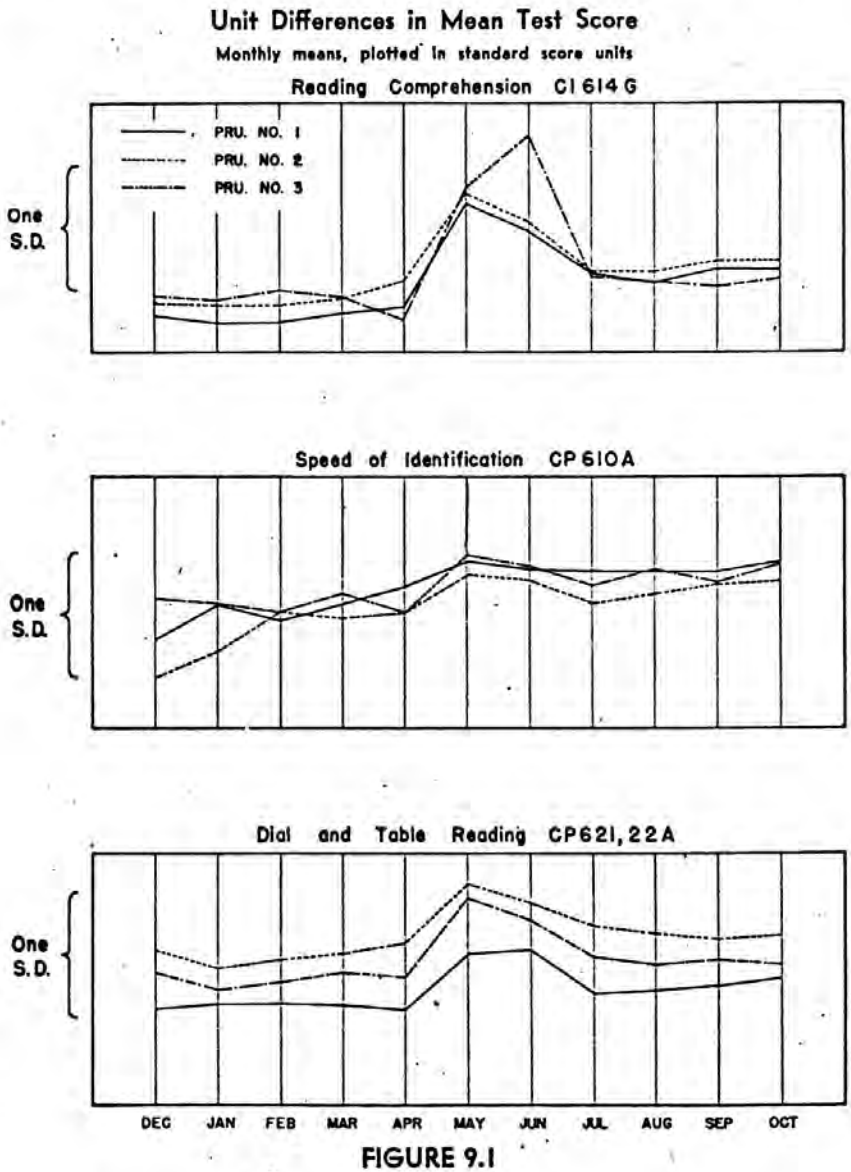


FIGURE 9.1

not on others was suspect with regard to the administration of that particular test. Several local divergences from standard procedure of scoring, timing, or administration were located through this type of systematic audit.

Sample results from this form of systematic auditing procedure are shown in figure 9.1. This figure shows monthly means at the three Psychological Research Units over a period of 11 months for the Reading Comprehension Test (CP614G), the Speed of Identification Test (CP610A), and the Dial and Table Reading Test (CP621, 22A). The Reading Comprehension Test results show little consistency in unit differences, but show a striking population shift in May 1943. This coincided with the introduction of the AAF College Training Program, and was due primarily to the policy of accelerating through the colleges those men who stood highest in the Educational Examination (AC20A) which was used as a screening device. Speed of Identification is another test for which unit differences are small and relatively inconsistent, though scores at PRU No. 2 tended to run low. In contrast, Dial and Table Reading shows relatively large differences between units, and differences which persisted without a single reversal for the period studied. It appears almost certain that procedures for administering this test were not the same at all units. The differences in this test are believed to have arisen from differences in the degree to which the basic instructions for the test were supplemented locally at the individual testing units. This test was one which involved rather complex instructions and in which supplementation of the standard instructions often seemed necessary. At the time that the testing shown in figure 9.1 was carried out, the backlog of statistical work so delayed these analyses that no corrective action was possible.

APPARATUS VARIANCE

In apparatus testing, it was always true that a number of different copies of each apparatus test were in use. The typical testing unit had two or three "lines" of apparatus tests. Each "line" contained four copies of each apparatus test in the battery, usually with a single control and cycling mechanism. Men were tested in squads of four, each examiner testing four men in a room at one time. As a result, each man was tested on a particular copy of each apparatus. Because of the use of a large number of copies of apparatus tests, a good deal of interest centered on the question of the comparability of the different copies and on establishing suitable procedures to guarantee comparable scores for an individual no matter on which copy of the apparatus he happened to be tested. The administration of complex apparatus tests on a large scale by the Aviation Psychology Program was

quite a novel undertaking. There has probably never been a time when apparatus tests were administered in as many different places and to as many different people under conditions which were as nearly uniform and standard for all individuals.

At a relatively early stage in the research program associated with classification testing, data were assembled and analyzed by copy of the test apparatus to determine whether the specific copy used was a significant source of variance in resulting test score. Analysis of variance and similar techniques applied to the distribution of raw scores showed quite clearly that for at least some of the apparatus tests the particular copy was a real source of variation. Apparatus differences were greater in those tests than could reasonably be attributed to sampling fluctuations. The amount of apparatus variance differed quite a bit for different ones of the apparatus tests. A relatively simple pegboard test requiring the subject to turn square pegs as rapidly as possible (Finger Dexterity, CM 116A) showed relatively little variation from one copy to another. It was apparently possible to construct the relatively simple equipment required for this test with sufficient uniformity so that differences from one copy to another were hardly a source of variance in test performance. Of the other classification tests, the Complex Coordination Test (CM 701A) showed probably the largest acount of variation from one copy of the apparatus to another. The variations in the earlier copies of the Complex Coordination Test were in part due to the complexity of the instrument and in part due to the fact that different copies had been made at different times with somewhat different detailed specifications. Other tests in addition to the Complex Coordination Test showed apparatus differences which were clearly statistically significant. These differences arose in spite of all the administrative provisions which had been made to achieve uniformity of testing conditions. Administrative provisions covered both procedures for test administration and routines for apparatus inspection and maintenance.

It is possible for certain sources of variation to be statistically significant in the sense that they could not reasonably have arisen by chance and yet for them not to be practically significant. It becomes appropriate, therefore, to inquire whether the obtained variation among copies of an apparatus are of practical significance as well as whether they meet standards of statistical significance. Practically significant variation for a test which is being used to predict success in some type of performance will be shown in reduction of the validity of that test. If the irrelevant variance introduced by apparatus differences is sufficient to bring about an appreciable reduction in the validity coefficient for the test, it may be considered to be of practical significance.

Unless some appreciable reduction in test validity results, the irrelevant variance cannot be thought to have practical significance. Formulae were developed[1] to estimate the reduction in test validity resulting from variance between copies of the test. The formula for the correlation between obtained score and "true" score, corrected for apparatus differences is

$$r = \sqrt{1 - \frac{\sigma_m^2 - \frac{n(\sigma_x^2 - \sigma_m^2)}{N}}{\sigma_x^2}} \qquad (1)$$

where σ_x^2 = the total variance of the distribution of scores
σ_m^2 = the variance of the apparatus means
n = the number of different copies of the apparatus
N = the number of cases tested

This formula allows for the variation among apparatus means which could have been expected from sampling alone, and the decrease of the correlation below unity is due only to the variation in excess of that attributable to sampling.

Assuming all variance introduced by apparatus differences to be error variance, and consequently unrelated to the criterion, the correlation of a "true" score with the criterion can be estimated by the formula

$$r_{ct} = \frac{r_{cr}}{r_{rt}} \qquad (1)$$

where r_{ct} = correlation of true score with criterion
r_{cr} = correlation of obtained score with criterion
r_{rt} = correlation of obtained score with true score

It is of some interest to examine the results from an early set of data as an instance of the amount of reduction in test validity which could have been expected to result if no provisions had been made for controlling variations among different copies of each test.

Data were analyzed for scores of slightly over 2,000 men tested at PRU No. 1 during 1 week in the spring of 1943. Twelve copies of each test were in use at that time. Variance was analyzed for score on the Complex Coordination, Two-hand Coordination, and Rotary Pursuit Tests. The variance of the total distribution of scores and the variance of the means for the separate copies of each test were determined. Applying formula (1) above, correlations between raw scores and scores corrected for apparatus differences were determined. These were then applied to hypothetical raw score validities of .400 for each test

[1] These formulae were developed by Lt. Colonel A. P. Horst.

703315—47—11

to show what corrected score validity would be expected in each hypothetical case if corrections for apparatus differences were applied. The results were as follows:

	Correlation of raw and corrected score	*Hypothetical raw score validity*	*Estimated true score validity*
Complex Coordination	.929	.400	.431
Two-hand Coordination	.982	.400	.407
Rotary Pursuit	.992	.400	.403

In the case of each of the tests, a previous comparison of apparatus variance with residual variance had indicated apparatus differences to be significant in the statistical sense. The data which have just been presented, however, suggest that it was only in the case of the Complex Coordination Test that they were of a size to be practically important. The type of analysis carried out here represents a worthwhile practical check on the losses which are resulting from apparatus differences (or some similar type of extraneous variance) and the gain in test validity which could be anticipated from the elimination or correction of those differences. This formula and method of analysis may, of course, be applied to the evaluation of any extraneous source of variance.

The results in the previous paragraph have indicated that the loss in validity due to apparatus differences would have been real for at least certain of the apparatus tests, though it cannot be said to have been large. In the composite aptitude score, of course, the loss in validity would have been reduced because those tests constituted only a limited part of the total score. The results which have been reported above are, of course, those which were obtained when careful procedures of standardization of test administration and of apparatus maintenance were in effect. Apparatus differences would presumably have been a very much more serious factor if these standardized procedures had not been carried out with great care. Under the circumstances it can be said that some slight loss in validity might have been expected from apparatus differences if no further effort had been made to correct for them, but that this would have been a relatively small effect in comparison to the over-all validity of the test battery.

In view of the finding that real apparatus differences existed, further efforts were made to provide for them in testing procedures. These efforts took two directions, one mechanical and one statistical. On the one hand, studies were carried out to determine what features of the apparatus were associated with apparatus differences and to invent calibration procedures designed to reduce differences in those features. For example, a measuring instrument was designed to measure the effective size of the target button on the Two-hand Coordination Test. Al-

though size had been specified in the original construction of the apparatus, different copies were found to vary within small limits. Comparison of apparatus means with measures of effective target size indicated that a substantial part of the variation from one apparatus to another was associated with this difference in effective size of target button. Once this was determined, routine calibration procedures were instituted for measuring target size and adjusting each copy of the apparatus so that this factor was maintained more nearly constant from copy to copy of the test. Similar calibration procedures were developed for size of the contact points in the Complex Coordination Test and for stylus pressure in the Rotary Pursuit Test. These calibration procedures served to reduce the obtained differences between copies of an apparatus so that variance from this source was ultimately reduced below practically significant values.

The second approach toward control of apparatus variance was statistical. This took the form of maintaining regular cumulative records of score distributions for each copy of an apparatus test. During the early period of the classification program, when differences between copies of the test were still significant, the score distributions were used as a basis for preparing separate standard-score conversion tables for each apparatus. Those standard-score conversion tables made possible the assigning of scores to an individual which took account of the characteristics of the apparatus on which he was tested. It must be remembered, however, that the development of conversion tables based upon separate samples for each copy of the test produced some sampling variation in the separate conversion tables. Where apparatus differences were large the sampling variation could be expected to be considerably less than the apparatus variation. When apparatus differences had been reduced, sampling variations became as significant a factor as apparatus differences and any gain from separate conversion tables was then lost. At this point, separate conversion tables were dispensed with.

The separate statistics on each copy of an apparatus always served as a basis for detecting apparatus malfunction. A continuous running record was maintained for successive groups of subjects. This record was used during the earlier period of the classification program both as a basis for revising conversion tables and as a basis for detecting the need for overhaul of a particular piece of apparatus. During the latter part of the classification program, records of successive hundreds of cases tested with a particular copy of an apparatus served as a basis for indicating need for special maintenance and also as a basis for removing a piece of apparatus from a testing line and substituting the spare copy if serious malfunction seemed to have

developed. The copy which had been giving aberrant scores was then returned to the School of Aviation Medicine for intensive mechanical overhaul.

EXAMINER VARIANCE

Administration of individual apparatus tests was carried out by a large number of different examiners and it was a matter of some concern whether the resulting test scores were to any considerable extent a function of the particular examiner who did the testing. Here again, every possible administrative precaution was taken to minimize variance from this source. The administration procedures for the different tests were specified in great detail. This included not only the instructions, which were prepared in standard form, memorized by the examiners and administered verbatim, but also details concerning the exact amount and type of demonstration and preliminary practice which was to be given for each of the separate tests. The statistical studies of variance between examiners refer, therefore, to the variance which remained after intensive precautions had been taken to reduce it to a minimum.

Studies of examiner variance showed somewhat divergent results in groups studied at different times and at different Units. This could well occur if the level of training and standardization of examiners varied from unit to unit and from time to time. Results from three separate representative studies are presented in table 9.1. In two of these studies, examiner variance was sta-

TABLE 9.1.—*Statistical Significance of Examiner Differences.*

Ratio of Between-Examiner to Within-Examiner Variance (F-Ratio) for Apparatus Tests in Three Samples.

Test	F-Ratio		
	PRU #2	MPEU #10	MPEU #7
Finger Dexterity	1.30	3.02	9.12
Rotary Pursuit	1.00	1.94	4.97
Discrimination Reaction Time	1.29	1.08	5.29
Two-hand Coordination	2.40	1.05	2.67
Complex Coordination	2.06	2.49	4.67
Rudder Control	....	1.21	3.47
Aiming Stress	1.77	....	....
F-Ratio for .05 Level of Significance	1.64	1.90	1.64
F-Ratio for .01 Level of Significance	1.99	2.46	1.99
Number of Subjects (Approximate)	900	500	5000
Number of Examiners	18	10	18

tistically significant, but just barely so, for three of six tests. In the other case, examiner variance met tests of significance for all tests. The tests for which the largest differences were found varied a good deal from study to study. The difference in size of F-ratios between the last study and the first two is primarily a function of the number of cases included in the study. The results from the last study indicate that when a large group is studied, statistically real examiner differences are found. It

was found, incidentally, that these differences were due to two or three aberrant examiners.

The practical significance of the above differences is a further question. Formula (1) on page 124 has been applied to the data for the Finger Dexterity Test. This is the test which showed the largest examiner differences in the MPEU No. 10 and MPEU No. 7 samples. As applied to these two sets of data, the correlations between obtained score and true score, freed of the influence of examiner differences, are .983 and .989 respectively. This means that very little attenuation of test validity resulted in these cases from examiner differences. Some of the instances of examiner difference appear to be statistically real, but it can be doubted that even these were of practical importance. With the level of standardization that was maintained in the Aviation Psychology Program, no appreciable attenuation of validity appears to have resulted from examiner differences.

VARIANCE ASSOCIATED WITH TIME OF DAY

The time of day at which tests were administered was another factor studied as a possible source of variance in psychomotor test scores. An initial study of this factor produced very disturbing findings in that very striking variation was discovered associated with time of day. The variance was significant by all routine tests of statistical significance and no artifacts were discovered which could reasonably have produced the differences which were observed. The above finding led to immediate repetition of the analysis with groups of data from several different units. The initial results were not confirmed. In the subsequent studies, time of day appeared not to be a significant systematic factor influencing apparatus test scores. No adequate rationalization of the difference between the initial study and the other studies which followed it has been developed. The results of five studies are summarized in table 9.2.

OTHER SOURCES OF VARIANCE

In addition to the factors considered in the previous sections, some attention was devoted to such factors as location in the examining room and newness of test booklets. Studies of position within the group test room were carried out both for printed tests and for motion picture tests. In the case of printed tests, one hypothesis held that men farther from the test administrator who was reading the instructions and directing the testing were at some disadvantage in following the procedures for the test. Another hypothesis held that, particularly in the case of certain highly speeded tests, men in the far corners of the room had a certain advantage in that they had a better opportunity to work

TABLE 9.2.—*A summary of studies of time of day and psychomotor test scores*

Unit	*N (approx.)*	*Starting testing times*	*Stat. analysis*	*Results*	*Remarks*
9 1st Study	2000 12-27 Nov. 1943	0745-1700	Times Grouped 0800-0900 1000-1500 1600-1700 Mean raw scores of each group compared to means of every other group.	8 of 12 CRs significant (P<1%). Striking trends for all tests save 2 Hnd. Coord. Performance maximal for early afternoon group.	When times grouped 0800-0900 and 1000-1500, % disqualified from each form of training significantly greater in early group. Equivalence of samples checked by comparing group test scores of members of 0800-0900 group with 1000-1500. No significant differences.
5	2500 10-15 Jan. 1944	0700-1800	Anal. of Var. Standard Score Means (15 min. groups).	Chance variations for all tests save Fing. Dext. (P<5%). The variances of time groups for Disc. React. Time are not homogeneous.	Some indication that the means of most of the tests are slightly lower in the morning.
7	4200 14-30 Dec. 1943	0745-1100 1200-1615 1800-1845 1945-2130	Anal. of Var. Standard Score Means (15 min. groups).	Chance variations for all tests save Rot. Pur. on which performance is significantly better (P<1%) for 0800 groups than others.	Rooms reported quite cold in early morning; suggested that low temperature might have affected speed of disk on Rot. Pur. Test.
8	950	0745-1130 1215-1615	Anal. of Var. Standard Score Means (Hour Groups).	Chance variations.	
9 2nd Study	1900 13 Jan. 1944 7 Feb. 1944	0745-0830 0945-1030 1215-1300 1415-1430 1445-1500	Times Grouped 0745-0830 0945-1030 1215-1300 1415-1500 Mean raw scores of each group compared to means of every other group.	4 of 36 CRs significant (P<1%). No striking trends. Slight tendency for performance to be minimal for early morning group, maximal between 1215 and 1300.	When times grouped 0745-1030 and 12-15-150), slight tendency for % disqualified from Bomb. or Nav. Training to be higher in morning group. % disqualified from Bomb. significantly greater. (P<1%). From Nav. (P<5%).

beyond the time limits without being detected. However, analysis of variance associated with sections of the testing room consistently failed to show any differences associated with that factor. In the case of motion picture tests, position in the testing room was particularly relevant in that each man was dependent upon the position of the motion picture screen for the stimulus patterns to which he responded. Studies of this factor are presented in more detail in Report No. 7 on motion picture tests. However, it may be said in general that for most of those tests rather wide tolerances in seat position were acceptable. In the case of most tests performance appeared to be relatively insensitive to position in the test room and the individual was apparently able to achieve a fairly high degree of size and shape constancy from whatever position he viewed the screen.

Wear and tear on booklets was a matter of some concern, particularly for certain of the perceptual tests in which rather fine discrimination of the material presented to the subject was required. Studies were made comparing performance of groups using new booklets and of groups using booklets which were so tattered that they were about ready to be salvaged as waste paper. No differences were found associated with conditions of the test booklet which approached statistical significance.

SUMMARY

In summary, it may be stated that insofar as it was possible to test such factors, results indicated that the conditions of testing within the Aviation Psychology Program were such as to keep irrelevant sources of variance within rather modest limits. Apparatus differences were real, and in certain instances large enough to be important. Both mechanical and statistical procedures were used to control these differences. As improved apparatus designs and calibration procedures were developed, statistical controls were used only to detect apparatus in need of maintenance or replacement. Other sources of variance were not shown to be of practical significance in the testing situation. The relatively small amount of variance attributed to apparatus, examiner, and other similar factors bears witness to the degree of standardization of procedures which was achieved in a systematic and carefully-controlled program.

CHAPTER TEN

Training Experiments

INTRODUCTION

Although the early focus of effort in the Aviation Psychology Program was on selection and classification of personnel, as the war progressed more and more attention was devoted to problems of training procedures. The general pattern appropriate for research on training procedures in the AAF conformed to the traditional pattern of the learning experiment. It was necessary that two or more alternative sequences of training activities be defined in terms of some hypothesis concerning efficient procedures for training; that groups be set up, preferably equated on relevant background characteristics, and then trained by each of the methods proposed; that the relative proficiencies of the groups be evaluated after completion of the sequence of training; and that appropriate statistical tests be applied to determine whether the observed differences in performance exceeded those that could be anticipated on the basis of sampling fluctuations. In abstract outline the pattern was clear and followed a thoroughly standard and well-known course. In actual practice, however, a variety of problems beset the research worker as he tried to make an actual field study conform to the specified pattern. The following sections will discuss the three areas of (a) definition of the training problem, (b) practical administrative problems in training research, and (c) criteria for use in training experiments.

THE DEFINITION OF THE PROBLEM IN TRAINING RESEARCH

In research on selection of personnel the problem involved was relatively unitary and clearly defined. All research efforts centered around the basic unifying problem of getting the most accurate prediction of an appropriate criterion of proficiency in the task in question. In the case of training research it might have been possible to specify verbally a similarly unifying central problem, that of achieving a maximum increase in the proficiency of personnel assigned for training. However, the separate studies which work toward the achievement of the general goal seem

to have much less of a unifying theme in the latter case than in the former. Training is a long involved process and almost every step or feature of it may be tinkered with by the research worker or by experimentally inclined training personnel. Each experiment may be planned and carried out to a considerable extent in independence of other studies. The cumulative result of a program of research is likely, therefore, to be a number of findings running off in a variety of directions. The problem of unifying and integrating research in this field is a very real one. A first and major concern for the psychologist, therefore, was effective definition of the problems for investigation.

Definition of psychological research problems in aircrew training involves two aspects. The first is the discovery of significant and testable hypotheses for alternate methods of training. The second is the definition of each of these hypotheses by a set of practical training operations. Hypotheses to be tested emerge in part from the general psychological literature on learning, in part from the practical hunches of training personnel.

On the one hand, the psychologist tends immediately to bring over and try to apply to aircrew training the findings of laboratory and classroom experimentation on knowledge of results, distribution of practice, transfer of training, and the like. Thus, in studying aircraft recognition training, it appeared to aviation psychologists that the factor of active response and immediate reinforcement of correct responses was of central importance. A little experiment was set up which demonstrated the importance of this factor. Again, in gunnery training aviation psychologists were very interested in the development of the Firing Error Indicator, a device which provided immediate knowledge of the direction of error in air-to-air firing. A chief reason for the unsatisfactory progress in aerial firing was thought to be the fact that while he was firing, the gunner received no information as to the amount and direction of his errors. Unfortunately, the engineering problems in the development of the Firing Error Indicator was never completely solved, so that the device was not available for experimental evaluation as an aid to training.

On the other hand, practical problems are continually arising within the training situation which require experimental study if an accurate and unbiased basis is to be provided for official decision and action. Aircraft recognition training had been built around the use of very brief "flash" exposures. Training personnel needed information as to whether these procedures were resulting in a maximum rate of learning, and so several experiments were run to check this point. A training aids officer had developed a special sight to be used in "skeet" training by fighter

pilots. Data were analyzed to determine whether training with this new type of sight gave more transfer to aerial firing than training with the standard sight. In navigation training, a very elaborate ground trainer had been developed to simulate dead-reckoning navigation missions. Training personnel were concerned to know whether use of this device improved navigational performance in the air, so a study was conducted to determine the effectiveness of this device. Personnel in charge of bombardier training desired to know how much aerial training was needed to bring student personnel to maximum accuracy in bombing, so a study was planned to plot the learning curve for aerial bombing.

From whichever source the problem or hypothesis comes, but especially when it comes from the psychologist, a further problem is involved in translating the general hypothesis into a specific set of training operations. That is, the general interest of the psychologist in transfer of training might be expressed more specifically as an inquiry concerning whether experience on a particular gunnery synthetic trainer will improve subsequent skill in air-to-air firing under simulated combat conditions. This hypothesis still requires a great deal of further detailed specification in terms of amount and kind of experience on the training device, personnel to be studied, and criteria to be used. Specification of a set of practical operating training procedures which at the same time provides an adequate test for the general hypothesis being studied and compares the most meaningful practical alternatives for a training program, is a difficult problem. For example, in the study of the dead-reckoning navigation trainer referred to in the previous paragraph, it was necessary to decide just what use of the trainer was to be evaluated. In actual fact, what was compared was use of the ground trainer in place of classroom work for several "ground problems." These problems were tasks similar to that of maintaining a navigational log in flight, except that instrument readings were synthetic and given on the ground—on the blackboard in the case of classroom instruction, or instrument dials in the case of the trainer. In the experiment, the trainer replaced the classroom as the locale for the ground problems which were a part of the standard curriculum of instruction. It is clear that the results of the experiment evaluate only *this* use of the trainer, and not other possible uses. That is, the trainer might have had special value as a device for supplementary instruction even though it was not found to have any as a substitute for the usual type and amount of classroom ground problem. In any event, the experimental results are a function of the specific definition of the use of the training device.

The necessity of planning the comparison of alternative training procedures in such a way that the test will be experimentally clear and practically meaningful calls for close cooperation of professional psychological personnel on the one hand and practical training personnel on the other. It should be facilitated by providing that the psychologist receive first-hand experience in aircrew training. At the same time, it calls for extensive indoctrination of line personnel in the research point of view toward training problems.

ADMINISTRATIVE PROBLEMS

Even when it had been possible to define a training research problem with a satisfactory degree of precision, the practical problems involved in setting up a training experiment in the official Army situation were very real. The Army training stations were engaged primarily in a large scale training enterprise and not in a research program. When personnel numbering in the thousands were being pushed through training at a maximum rate in order to meet commitments for combat operations, it was not easy to interfere with the course of this training in order to set up ideal conditions for research work. The typical training station was a large, complex, highly integrated organization, all the parts of which had to work together smoothly if training was to proceed. Any considerable modification of training procedures for research purposes was likely to produce eddies of disturbance in the normal smooth flow of training operations in other groups in the station. A special time schedule for flight missions for the experimental groups would have meant that the flight schedules for other groups had to be rearranged. Special planes for the experiment would have meant withdrawing these temporarily from the supply at the station, which might already be limited. Maintaining the best calibration of instruments in experimental planes would have meant an additional burden on an already overtaxed maintenance staff. Allocation of a special group of officers to ride extra checks upon the experimental students would have put demands on the limited checking staff at a given station which would have been beyond all reason.

As the war moved towards its successful conclusion, with a consequent lessening of pressure on training personnel, as aviation psychologists became better established in training research, and as training research projects were backed up by stronger directives from higher headquarters, it became possible to do more and more in the way of setting up special conditions for research projects. Thus, in 1945 an experimental study of the learning curve for aerial bombing was directed by Headquarters AAF, and for this project it was possible to obtain jurisdiction

over the personnel selected to be trained, the personnel to carry out the training, the planes used for the aerial bombing and the pilots who flew them, the bombsights and the schedule for their maintenance, the daily schedule of bombing, and the conditions of altitude, bomb run, evasive action and the like under which each bomb was to be dropped. There was, unfortunately, one feature over which it was not possible to obtain control, which rather thoroughly upset the plans for this study. This was the end of the war with Japan.

Another element of difficulty was that research activities had to be carried out through the medium of standard operating personnel with a decidedly limited background of research experience and research interest. Thus, when a standard flight test was administered to evaluate the effect of an additional five weeks of flight training for certain classes which were held over for that additional period, the tests were administered by several hundred check pilots who had only a limited session of indoctrination in the nature, purpose, and techniques of administration of such a standard check. Again, the evaluation of the navigation ground trainer was based upon the instruction provided by standard instructional personnel and upon test flights flown by regular service pilots and evaluated by regular staff navigators. The degree to which standardization of training procedures and control of experimental conditions could be maintained while depending upon large groups of personnel of this type is a matter of question.

Problems of equating groups presented particular administrative difficulties in military research, in that groups were already set up in administrative units and any interference with those units raised serious administrative problems. Experience showed that successive groups which had been set up by administrative procedures did not represent random samples from the same population, so that one could not rely upon the equivalence of the groups as they were already found to exist. In specially planned studies, such as those which we have already mentioned dealing with the navigation ground trainer and with the learning curve for aerial bombing, it was possible to select and equate the experimental groups. However, it was always bothersome and sometimes impossible to interfere with personnel assignments so as to achieve this end.

The above problems are not ones which have any particular theoretical significance. They did not unduly tax the research workers' intellectual abilities in determining what should ideally be done to take care of them. They did, however, pose very severe problems for the research workers' personal tact, ingenuity and administrative skill.

THE CRITERION IN TRAINING RESEARCH

In practice the most difficult technical problems in training research were concerned with the establishment of adequate techniques for measuring proficiency in the tasks for which training was being given. The general problems of criteria of proficiency have been discussed in an earlier chapter of this report. These problems were basically the same whether the criterion measures were required for aptitude test valuation or for evaluation of results of training experiments. However, certain specific points are worthy of mention in connection with the establishment of proficiency measures for training research.

Since training studies are carried out on a group basis and it is group results in which one is interested, high reliability in the criterion is not a critical requirement in these studies. Since, however, a comparison is being made of the systematic effect of two or more distinct procedures, it is imperative that the criterion measure be unbiased. That is, there must be no possibility that the criterion test is being administered under conditions which permit one group to have an advantage relative to the other. This tends to be less critical in the case of selection test validation, because biasing factors are likely to be randomized with respect to the factor being studied (aptitude test score). Thus, different schools, different instructors, different check pilots are likely to get essentially random samples with regard to aptitude test score. Any bias due to school, instructor, or check pilot will then be spread out so it affects all levels equally, and will become an attenuating rather than a systematically biasing factor. In the case of training experiments, however, the two groups being compared are likely to be discrete groups, trained at a different time or place, or at least to be distinguished and identified by training personnel. In this case any biases associated with a particular school, a particular flight, a particular group of instructors, or the attitude of the group of instructors toward the two groups are likely to affect the groups differentially. The biases may become systematic rather than random. When this happens, the validity of the experimental results is lost. Measures of proficiency based upon subjective ratings and evaluations are particularly suspect from the point of bias. Changes from time to time and place to place are the rule rather than the exception. Furthermore, in an experiment being carried on at a given time and place it is entirely possible that raters may be biased in favor of one rather than the other of the methods under study and that they may prejudice their ratings accordingly. One is led to conclude that it is particularly in training experiments that complete objectivity is needed in the criterion measure.

The question may be raised whether the same types of measures of proficiency are appropriate for evaluating the results of training as are appropriate for subsequently validating aptitude measures. Studies of objective flight items as measures of pilot proficiency suggested that those which were predictive of pass-fail in training were rather different from those which discriminated groups with different amounts of training. It would seem that proficiency measures for evaluating training procedures would need to be related a good deal more specifically to the particular knowledges and skills included in the training program than would measures used as a criterion for the evaluation of aptitude tests.

APPENDIX A

The AAF Training Command Correlation Chart

The AAF Training Command correlation chart has been designed to take advantage of economies that can be effected through the use of computing machines and to provide a complete series of checks on the computation of all constants needed in finding the coefficient of correlation. The chart presents no particular advantages in insuring the accuracy of the original scatter diagram or in computing the coefficient of correlation after the required constants have been obtained.

Preparation of the Scatter Diagram. Spaces are provided on the left-hand margin and the upper margin for indicating the step intervals. A maximum of 21 steps may be used in either dimension. In order to avoid the use of negative quantities, only the positive quadrant is used, the x-origin being the midpoint of the step at the extreme left and the y-origin being the midpoint of the lowest step. The dx's and dy's, which are indicated in spaces adjacent to step intervals, are used only in the computation of the cross products. The column and row in which these values are zero indicate the arbitrary origin in x and y respectively. The scatter diagram is prepared in the usual fashion. Two methods of checking the scatter diagram are feasible: the preparation of a duplicate diagram from the original data by another clerk, or the preparation of distributions of the two variables using the same step intervals as employed on the correlation chart. If the latter method is used, the distributions should be cumulated toward the lowest step and compared with the cumulative distributions found on the chart. The preparation of the scatter diagram may be facilitated by writing the x steps on a strip of squared paper having the same size squares as employed on the chart. If the step containing the y score is found first, this strip may be placed below that step and the entries used as a guide in locating the proper cell in the x column in which the tally is to be made.

Finding the Cumulative Frequencies. In using this chart cumulative frequencies, rather than frequencies, are found for both

NUMERICAL EXAMPLE

Showing Certain Steps in the Computation of the Coefficient of Correlation from the AAF Training Command Correlation Chart. Artificial Data. N=20.

X → / Y ↓ (dx / dy)	8/9	10/11	12/13	14/15	16/17	m	Cfy	m′	C Σ $d_x f_{xy}$
50–59	4					1	7	1	9	4
40–49	3				2	1	1	5	5	7	15
30–39	2		1	2	2	2		3	12	5	27
20–29	1		2	3	1			1	18	3	32
10–19	0		1		1			(N)→	20	1	34
		(N)	1	3	5	7	m			Σ x′
		20	16	11	5	2	Cfx			
		1	3	5	7	9	m′			
Σ y′		36	32	25	14	7	C Σ $d_y f_{xy}$			

N 20 (v)

Σ mCfy = (Σ y′²) 86 (v)

Σ Cfy (Σ y′) 36 (v)

Σ y′ 36

Σ m′Cfy 178

$$B = \frac{424}{N\Sigma y'^2 - (\Sigma y')^2}$$

$\sqrt{B}$ 20.591

i_y 10

$$\frac{i_y \sqrt{B}}{N} = 10.296 = a_y$$

rows and columns. All work in either the x or y variable starts as far from the origin as possible and is carried toward the origin of that variable. Starting with the first row in which there are tallies, the cell frequencies are added into the adding machine or calculator and the total of the cell frequencies in that row is entered in the column headed Cfy. This sum is not cleared from the machine but is added to the cell frequencies in the row below to form the cumulative frequency for that row. If a row has no tallies, the Cfy is the same as that of the preceding row. The Cfy of the bottom row is necessarily N, the number of cases. The procedure used is readily apparent from an inspection of the numerical example in which the successive Cfy's are 1, 5, 12, 18, and 20.

By a process exactly analogous, the cumulative frequencies in x are found. Tallies in the column farthest to the right are added to find the first Cfx. Without clearing the machine, the tallies in the column to the left are added to find the next Cfy so on across to the column containing the x-origin, the Cfx of which is N. Thus, in the numerical example, the Cfx's are 2, 5, 11, 16, and 20. It is to be noted that N is determined twice.

The Computation of $\Sigma y'$. To obtain $\Sigma y'$, the Cfy's are added, excluding the entry in the step which contains the assumed mean. It is to be noted that the assumed mean is at the x- or y-origin, denoted by zeros on the chart. This method of computing the sum of the deviations in terms of step intervals from the arbitrary origin takes advantage of the principle that the sum of a series of cumulative frequencies is equal to the sum of the products of each frequency times its deviation from the origin in terms of step intervals. This fact is easily noted from Algebraic Example I.

ALGEBRAIC EXAMPLE I

f	d	df	Cf
a	n	na	a
.	.	.	.
.	.	.	.
.	.	.	.
h	3	3h	a + ... + h
i	2	2i	a + ... + h + i
j	1	j	a + ... + h + i + j
		ΣCf =	na + ... + 3h + 2i + j

The frequencies are indicated in the column headed f and are a ... h, i, j. The column headed d gives the deviations in terms of step intervals from the arbitrary origin. The column headed df gives the products of the deviations in step intervals as obtained in the ordinary multiplicative method of computing the mean from an arbitrary origin. The column headed Cf gives the cumu-

lative frequencies. Since there are n of these cumulative frequencies, in all which a is represented, na will be represented in the ΣCf. It is readily apparent that irrespective of the number of terms or the values of the frequencies the sums of the two columns df and Cf are identical. In summing the cumulative frequencies to obtain $\Sigma x'$ or $\Sigma y'$ care must be taken not to include N in the step containing the origin.

Computation of $\Sigma y'^2$. To compute $\Sigma y'^2$, the Cfy's are multiplied by the successive odd numbers beginning with unity in the step above the one which contains the assumed mean. The sum of these products is $\Sigma y'^2$. This method of computing the sum of the squares of the deviations in terms of step intervals from the arbitrary origin is an application of the fact that the sum of a series of odd numbers beginning with unity is equal to n^2 when n is the number of terms in the series. The actual employment of this principle is indicated in Algebraic Example II.

ALGEBRAIC EXAMPLE II

f	d^2	Cf	m	$mCfy$
$\overline{a}$	$\overline{n^2}$	$\overline{a}$	$2\overline{n-1}$	$(\overline{2n-1})a$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
h	9	$a + \ldots + h$	5	$5a + \ldots + 5h$
i	4	$a + \ldots + h + i$	3	$3a + \ldots + 3h + 3i$
j	1	$a + \ldots + h + i + j$	1	$a + \ldots + h + i + j$
				$\Sigma mCf = n^2a + \ldots + 9h + 4i + j$

The successive odd numbers are denoted as m, or the multiplying factors. It will be seen that $\Sigma mCf = \Sigma d^2f$. Again the cumulative frequency in the step containing the assumed mean is ignored.

Numerical Computation of $\Sigma y'$ *and* $\Sigma y'^2$. When a key-driven adding machine is used to compute $\Sigma y'$ and $\Sigma y'^2$, two series of operations are performed: the summing of the Cf's and the summing of the products of each Cfy with its corresponding m. When a calculating machine is used, the two quantities are found in one series of operations. The m's are placed in the keyboard and multiplied by the corresponding Cfy's. The accumulation of the multipliers is ΣCfy or $\Sigma y'$ and the accumulation of products in the two product dials is $\Sigma mCfy$ or $\Sigma y'^2$. In the numerical example ΣCfy is 36 and $\Sigma mCfy$ is 86. $\Sigma x'$ and $\Sigma x'^2$ are computed similarly and the four quantities are entered in the appropriate spaces under "Computations." It is to be noted that there are two spaces for entering $\Sigma x'$ and two spaces for $\Sigma y'$. Entries are to be made in both places in connection with Charlier's check on the sums of squares.

AFTRC Form No. 26
1 March 1945
Auth: T.C. Ltr. 25-3

ARMY AIR FORCES TRAINING COMMAND

CORRELATION CHART

Incl No 1 to TC Ltr. 248 (actual size 19½ x 28 in.)

Y \ X; dx / dy	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	m	Cfy	m'	$Cd_x f_{xy}$
20																						39		41	
19																						37		39	
18																						35		37	
17																						33		35	
16																						31		33	
15																						29		31	
14																						27		29	
13																						25		27	
12																						23		25	
11																						21		23	
10																						19		21	
9																						17		19	
8																						15		17	
7																						13		15	
6																						11		13	
5																						9		11	
4																						7		9	
3																						5		7	
2																						3		5	
1																						1		3	
0																						(N) →		1	
	(N) ↓	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	37	39	m			$\Sigma x'$
																						Cfx			
	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	37	39	41	m'			
$\Sigma y'$																						$C\Sigma d_y f_{xy}$			

COMPUTATIONS

N _____ ( )

$\frac{\Sigma mCfx}{(\Sigma x'^2)}$ _____ ( )

$\frac{\Sigma Cfx}{(\Sigma x')}$ _____ ( )

$\Sigma x'$ _____

$\Sigma m'Cfx$ _____

A $N\Sigma x'^2 - (\Sigma x')^2$ _____

$\sqrt{A}$ _____

i_x ☐

$\frac{i_x}{N}\sqrt{A}$ _____ $= \sigma x$

i

N _____ ( )

$\frac{\Sigma mCfy}{(\Sigma y'^2)}$ _____ ( )

$\frac{\Sigma Cfy}{(\Sigma y')}$ _____ ( )

$\Sigma y'$ _____

$\Sigma m'Cfy$ _____

B $N\Sigma y'^2 - (\Sigma y')^2$ _____

$\sqrt{B}$ _____

i_y ☐

$\frac{i_y}{N}\sqrt{B}$ _____ $= \sigma y$

IDENTIFICATION DATA

Organization _____

Project No. _____

Study _____

Variable x _____

Variable y _____

Computer _____

Verified by _____

Date _____

$\Sigma C\Sigma d_x f_{xy}{}^2$
$\Sigma C\Sigma d_y f_{xy}$
$(\Sigma x'y')$ _____ ( )

C $N\Sigma x'y' - \Sigma x'\Sigma y'$ _____

$C/\sqrt{A}\sqrt{B}$ _____ $= r_{xy}$

$\frac{\quad}{(M'_x)} + \frac{\quad}{\left(\frac{i_x \Sigma x'}{N}\right)} = \frac{\quad}{} = M_x$

$\frac{\quad}{(M'_y)} + \frac{\quad}{\left(\frac{i_y \Sigma y'}{N}\right)} = \frac{\quad}{} = M_y$

013145-1

(Facing p. 150)

Charlier's Check. The basic formula for Charlier's check is readily derived as follows:

$$(y' + 1)^2 = y'^2 + 2y' + 1$$

$$\text{Summing, } \Sigma(y' + 1)^2 = \Sigma y'^2 + 2\Sigma y' + N$$

This is to say, if we drop the assumed mean one step interval and thereby increase the value of all deviations by 1, the new sum of squares will be equal to the old sum of squares plus twice the sum of the deviations from the old origin plus N. The successive odd numbers denoted by m' are designed for use with Charlier's check. The procedure used in determining ΣCfy and $\Sigma mCfy$ is repeated and the sum of the products of the m's and the Cfy's entered under $\Sigma m'Cfy$. When it appears that this value is equal to the sum of the four entries above it, a check is obtained on the computation of $\Sigma y'^2$. In the numerical example, $20 + 86 + 36 + 36 = 178$.

Computation of $\Sigma x'y'$ (Working from the Columns). $\Sigma x'y'$ is computed twice. Working from the columns, each cell frequency (designated as f_{xy}) is multiplied by the corresponding d_y and the sum of these products from the column at the right is entered in the space provided in the row labeled $C\Sigma d_y f_{xy}$. Without clearing this sum from the machine, the cell frequencies in the next column nearer the origin are similarly multiplied by the d_y's and the cumulative sum entered in the space provided. The work is carried through the column which includes the x-origin. The last entry is $\Sigma y'$ but this entry is used only as a check upon the previous computation of $\Sigma y'$. The sum of all the entries in this bottom row of the diagram, excluding the entry in the column containing the x-origin, is $\Sigma C\Sigma d_y f_{xy}$ or $\Sigma x'y'$. This quantity is entered in the appropriate space under "Computations." In the numerical example, the successive entries are $(1 \times 4) + (1 \times 3) = 7$; $7 + (1 \times 3) + (2 \times 2) = 14$; $14 + (2 \times 3) + (2 \times 2) + (1 \times 1) = 25$; $25 + (2 \times 2) + (3 \times 1) = 32$; $32 + (1 \times 2) + (2 \times 1) = 36$. $\Sigma y'$ is 36 both in this operation and in the operation involving the Cfy's. $\Sigma C\Sigma d_y f_{xy} = \Sigma x'y' = 32 + 25 + 14 + 7 = 78$. It is to be noted that the entries in the row containing the y-origin are not used in computing the cross products.

Computation of $\Sigma x'y'$ (Working from the Rows). The process of computing $\Sigma x'y'$ from the rows is exactly analogous to the process of computing $\Sigma x'y'$ from the columns. Beginning with the row farthest from the origin, work proceeds toward the origin in y. Each cell frequency (f_{xy}) is multiplied by its corresponding d_x and the product is entered in the extreme right-hand column of the chart headed by $C\Sigma d_x f_{yx}$. Without clearing the first entry from the machine, the entry for the next row is computed and entered in the allotted space. This continues through the row con-

taining the origin in y in which case the entry is $\Sigma x'$ and becomes a check on the previous computation of this figure. The sum of the entries in the column, excluding the final entry, is $\Sigma C\Sigma d_x f_{xy}$ or $\Sigma x'y'$. In the numerical example $(1 \times 4) = 4$; $4 + (2 \times 2) + (1 \times 3) + (1 \times 4) = 15$; $15 + (2 \times 1) + (2 \times 2) + (2 \times 3) = 27$; $27 + (3 \times 1) + (1 \times 2) = 32$; $32 + (1 \times 2) = 34$. 34 is the $\Sigma x'$ previously found and $4 + 15 + 27 + 32 = 78 = \Sigma x'y'$. In computing $\Sigma x'y'$ either from the rows or from the columns, it will be found useful to have a strip of cardboard with the d's indicated on it. Such strips may be cut from a copy of the chart and used repeatedly.

Algebraic Explanation of the Computing Principle in Finding $\Sigma x'y'$. The algebra in determining each $x'y'$ will be readily observed. Consider a tally in a cell with a y value of a and an x value of b. When working in the columns this tally takes on a value of a and, since cumulative sums are used, it appears in the $C\Sigma d_y f_{xy}$ row b times, and hence adds ab to the value of $\Sigma x'y'$. The same tally appears as b, a times in the $C\Sigma d_x f_{xy}$ column.

Computations. The computations leading to σ_y are illustrated in the numerical example and follow the familiar pattern. After N, $\Sigma y'^2$ and $\Sigma y'$ are found and checked, the quantity B, defined as $N\Sigma y'^2 - (\Sigma y')^2$, is obtained by use of a calculating machine and the square root of this quantity determined. Similarly, A is found for the x-variable. $N\Sigma x'y' - \Sigma x'\Sigma y'$ (denoted as C) divided by $\sqrt{A}\sqrt{B}$ is r_{xy}. The standard deviation of x is found by the formula $\frac{i_x}{N}\sqrt{A}$ and the mean of x is found from the formula $M_x = M_x' + \frac{i_x \Sigma x'}{N}$, in which i_x is the x-step interval and M_x' is the midpoint of the step containing the x-origin. The mean and sigma of y are found similarly. Work spaces for obtaining these statistics are included on the chart.

Recapitulation. Although this description of procedures is rather long, it will be found that operators can be trained to follow all steps quickly. The suggested routine is as follows:

a. Sum all the cell frequencies in the rows, obtaining the cumulative frequencies in y, the last entry being N.
b. Sum the frequencies in the columns, obtaining the cumulative frequencies in x, the last entry again being N.
c. Obtain $\Sigma y'$ by summing the cumulative frequencies down to, but not including, the step containing the assumed mean in y and $\Sigma y'^2$ by multiplying the m's by the corresponding Cfy's. Check the results through the use of the Charlier formula. Repeat the routine to obtain $\Sigma x'$ and to obtain and check $\Sigma x'^2$.

d. Compute $\Sigma x'y'$ by multiplying the cell frequencies in each column by the corresponding d_y, cumulating all results toward the origin. The last entry in the row labeled $C\Sigma d_y f_{xy}$ is $\Sigma y'$. The sum of the other entries is $\Sigma x'y'$. By a similar process in the rows, in which each cell frequency is multiplied by its corresponding d_x and the resulting sums are cumulated toward the y-origin, the $C\Sigma d_x f_{xy}$'s are found. In this way $\Sigma x'$ and $\Sigma x'y'$ are checked.

e. The coefficient of correlation is obtained by the usual formula.

$$r = \frac{N\Sigma x'y' - \Sigma x'\Sigma y'}{\sqrt{N\Sigma x'^2 - (\Sigma x')^2}\sqrt{N\Sigma y'^2 - (\Sigma y')^2}}$$

The means and sigmas of the two variables, if desired, may be obtained by the use of the usual formulas, which are indicated on the chart.

NOTE ON STEP INTERVAL AND THE ASSUMED MEAN

In fixing step intervals for handling psychological test data *when all scores are integral*, the true lower limit of the step is generally considered to be .5 of a unit below the integral lower limit actually written for the step. Thus, if a step interval is written 15-19, and all scores of 15 through 19 are tallied on this interval, the true lower limit is 14.5. To find the midpoint, one-half of the step interval should be added to the true lower limit. In this case the midpoint would be 14.5 + 2.5 or 17. As another example consider the case in which the intervals are written 0-9; 10-19; etc. The midpoint of the first interval is −.5 + 5.0 or 4.5; the midpoint of the second is 9.5 +5.0 or 14.5, etc. When the mean is to be computed, this rule should be followed in determining the assumed mean.

APPENDIX B

An Iteration Method for Determining Multiple Correlations and Regression Weights

Much of the procedure described here has been presented by Kelley and Salisbury,[1] and the present procedures are essentially an adaptation of their technique. No effort is made to provide here a complete or rigorous mathematical basis for the formulas and computational procedures, and the rationale is merely sketched in order to give an intuitive feeling for what is being done. Formulas for the various operations in this method are presented both in matrix notation and in the conventional scalar notation. The more compact matrix notation is presented on the left, and on the right the same operations are indicated in scalar notation for those to whom the matrix notation is unfamiliar.

The basic relationship upon which the iterative method depends is as follows:

$$\mathbf{r'_c} = \boldsymbol{\beta}'\mathbf{r} \qquad \text{or} \qquad r_{ic} = \sum_{j=1}^{n} \beta_j r_{ij} \tag{1}$$

That is, the correlation of a test i with the criterion is equal to the sum of the products of each test's beta weight and its correlation with test i.

The analysis starts with a square table of obtained correlations among tests in a battery and a column of empirically determined test validities. First a guess is made as to the beta weights for this set of data. Some of the considerations which enter into making a shrewd initial guess are considered later. To the set

[1] Kelley, T. L. and Salisbury, F. S. *loc. cit.* pp. 282 ff.

SAMPLE PROBLEM

Intercorrelations

Variable	*1*	*2*	*3*	*4*	*5*	*6*	*Pilot Validity*
1. Two-hand coord.	1.00	.27	.30	.48	.33	.23	.32
2. Fing. dext.	.27	1.00	.11	.35	.33	.26	.12
3. Rud. cont.	.30	.11	1.00	.32	.32	.07	.40
4. Comp. coord.	.48	.35	.32	1.00	.38	.36	.39
5. Rot. pur.	.33	.33	.32	.38	1.00	.18	.23
6. Disc. react. time	.23	.26	.07	.36	.18	1.00	.23

Calculation Sheet

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII
	$\bar{r}_c$	Wt_1	$\bar{r}_{c(1)}$	$\bar{r}_{c(2)}$	$\bar{r}_{c(3)}$	$\bar{r}_{c(4)}$	$\bar{r}_{c(5)}$	$\bar{r}_{c(6)}$	$\bar{r}_{c(7)}$	$\bar{r}_{c(8)}$	$\bar{r}_{c(9)}$	$\bar{r}_{c(10)}$	$\bar{r}_{c(11)}$	$\bar{r}_{c(12)}$	Wt_{12}	$\bar{r}_c$	$\bar{r}_{c(13)}$	Wt_{13}
1	.32	.10	.256	.279	.309	.298	.308	.310	.313	.323	.320	.317	.319	.315	.11	.315	.318	.11
2	.12	.00	.119	.145	.156	.116	.123	.126	.129	.132	.122	.121	.124	.120	−.05	.116	.119	−.05
3	.40	.20	.294	.301	.401	.397	.403	.404	.407	.410	.409	.399	.400	.397	.29	.396	.399	.29
4	.39	.20	.312	.348	.380	.366	.386	.390	.394	.399	.396	.393	.397	.387	.21	.385	.389	.21
5	.23	.00	.173	.191	.223	.210	.218	.220	.230	.233	.230	.227	.229	.225	.01	.224	.234	.02
6	.23	.00	.109	.209	.216	.206	.213	.223	.225	.227	.224	.223	.233	.229	.12	.230	.232	.12
			Var. #6	Var. #3	Var. #2	Var. #4	Var. #6	Var. #5	Var. #1	Var. #2	Var. #3	Var. #6	Var. #4			Var. #5		
Adjustment			+.10	+.10	−.04	+.02	+.01	+.01	+.01	−.01	−.01	+.01	−.01			+.01		

$$R = \frac{.2593}{\sqrt{.2590}} = .51$$

of guessed beta weights and to the obtained table of intercorrelations there corresponds some set of validities which satisfies equation (1). That is, if we designate the estimates of the betas $\tilde{\beta}$, we have

$$\tilde{r}_c = \tilde{\beta}' r \qquad \text{or} \qquad \tilde{r}_{ic} = \sum_{j=1}^{n} \beta_j r_{ij} \tag{2}$$

The estimated set of beta weights yields a set of validity coefficients, i. e., that set of validity coefficients for which this would be the exact set of beta weights.

The calculation of the $\tilde{r}_c$ values is simple and quite speedy if a Marchant or other good calculating machine is available. The intercorrelations are set up in a square matrix with unity for the diagonal terms. A calculating sheet is set up with lines spaced at the same distance apart as the vertical spacing in the table of intercorrelations, and with variables numbered to correspond to the variables in the correlation table. Column I of the computing sheet contains the empirical validity coefficients. Column II contains the initial guessed beta weights. Column III contains the $\tilde{r}_c$ values, and is obtained by placing Column II alongside each column of the correlation matrix in turn and getting the cumulative sum. This is shown in the illustrative example on pages 154 and 155.

Once an initial set of $\tilde{r}_c$ values has been obtained, the procedure becomes one of successive corrections to the beta weights one at a time until the $\tilde{r}_c$ values correspond to the empirical r_c's within a specified limit of accuracy. In general, one starts with the variable for which the discrepancy between r_c and $\tilde{r}_c$ is greatest, adjusts the beta weight by an amount which will approximately eliminate that discrepancy, and then computes a new set of adjusted $\tilde{r}_c$ values. (Column IV of example.) The procedure for making that adjustment is considered below. A second beta weight is then corrected and another new set of $\tilde{r}_c$ values obtained, and so forth. With practice a certain knack is developed in selecting variables to adjust and deciding upon the amount of adjustment to make. Adjustments are continued until the $\tilde{r}_c$ and r_c values are in sufficiently close agreement. In most work in the Aviation Psychology Program it was required that the two sets should agree exactly when rounded to two decimal places.

The first principle for adjusting beta weights is to adjust first the beta weight for which $|r_c - \tilde{r}_c|$ is greatest, and adjust it by the amount $r_c - \tilde{r}_c = d$. An adjustment of the weight for one variable will in general affect all the $\tilde{r}_c$ values. If we call the adjusted values $\tilde{r}_{c(2)}$ we have

$$\tilde{r}_{c(2)} = \tilde{r}_c + d_k r_k \qquad \text{or} \qquad \tilde{r}_{ic(2)} = \tilde{r}_{ic} + d_k r_{ik} \tag{3}$$

This can easily be seen if we expand the terms of equation (2) both for $\bar{r}_c$ and $r_{c(2)}$. If the adjustment d_k is a fairly small amount or a round figure such as .05, each correlation in column k of the correlation matrix can be multiplied by d_k mentally, the product subtracted mentally from the corresponding entry of Column III of the calculation sheet, and the difference entered in Column IV. Column IV then becomes the column $\bar{r}_{c(2)}$ of adjusted $\bar{r}_c$ values. A second adjustment can be made on Column IV in the same way, and so on. The beta weight next to be adjusted is always determined by comparing the column of $\bar{r}_c$ values resulting from the immediately preceding adjustment with the r_c column (Column I) on the calculating sheet, and noting the location of the greatest discrepancies. A check upon the accuracy of one's mental arithmetic, and upon the accumulation of rounding errors, is possible at any point by repeating the operation of formula (2) with the most recent approximation to the beta weights. (See Column XVI of sample problem.)

The composite correlation resulting from any set of weights may be computed quite simply. It is given by the following formula:

$$R = \frac{V'r_c}{\sqrt{V'rV}} \quad \text{or} \quad R = \frac{\sum_{i=1}^{n} V_i r_{ic}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} V_i V_j r_{ij}}} \qquad (4)$$

Where V signifies the weight attached to a particular variable. When the weights V correspond exactly to the regression weights, this formula simplifies to:

$$R = \sqrt{\beta' r_c} \quad \text{or} \quad R = \sqrt{\Sigma \beta_i r_{ic}} \qquad (5)$$

Using formula (4), it is possible to determine the correlation between any set of weighted scores and an additional criterion variable. This is frequently convenient in other problems in addition to the present one. In the present project, the formula can be used to yield a composite correlation at any particular stage in the approximation procedure, as well as at the end when the approximation has reached the desired standard of accuracy, at which point the composite correlation will approximate the multiple correlation resulting from true regression weights.

In actual computation, the numerator of (4) is the sum of products of a column of the latest set of weights, each times the corresponding validity coefficient in Column I. The expression under the square-root sign is the sum of products of weights times corresponding $\bar{r}_c$ values, i. e., times the validity coefficients produced by that set of weights.

Although it would be possible to start from uniform weights for all tests or from any other set of weights, a good deal of time can be saved if a close approximation to the correct weights is initially chosen. Kelley and Salisbury suggest starting by giving each test a weight one-half its validity coefficient. However, it is believed that with a little practice considerably more efficient skills in that regard can be developed. The following suggestions represent certain insights from working with the method.

a. If a set of weights is available from previous data or some other source, it can usually be used to advantage as a starting point. Thus, if a new set of pilot weights is being computed incorporating new validity data or based on new intercorrelations, one would ordinarily take the previous set of pilot weights as a first approximation.

b. When one is starting from scratch and is working with a substantial number of variables, it is usually sound to give about half of the tests, those having the lowest validities, a weight of zero to start with.

c. For the other tests, the weights should vary from about one-fourth to one-half of the validity coefficient. The highest fraction of the validity coefficient is used for the tests which have the highest validity coefficients or appear to have low correlations with other weighted tests.

There are one or two tricks in applying corrections to the initial weights also.

a. A purely routine procedure will require that each correction be made in the exact amount of the discrepancy between $\bar{r}_c$ and r_c. Experience shows that this procedure is frequently likely to result in overcorrection. If an inspection of the initial set of $\bar{r}_c$ values shows them to be either predominantly too high or too low, so that almost all of the indicated corrections are in the same direction, corrections should be made smaller than the amount $r_c - \bar{r}_c$. This is due to the fact that, when intercorrelations are largely positive, the corrections on different variables tend to supplement one another.

b. It is believed that time is saved for the whole operation by making most of the early corrections by convenient amounts such as .10 or .05.

A practical advantage of the present iterative method is that it makes it very simple to add any desired additional conditions to the set of weights one is computing, and then compute the most valid set of weights satisfying those conditions. The additional condition which was imposed in much of the work in the Aviation Psychology Program was that no weights should be negative. (In this case, weights were corrected down as far as zero, but no further correction was made.) It is also a simple matter to

drop out a test or group of tests (give them zero weights) and determine what the weights should then be for the rest of the tests. Any other desired conditions could be imposed in similar fashion.

The techniques which are described have been criticized as suffering from subjectivity in the determination of the order and amount of the corrections. It is possible that two computers might take the same set of data and come out with two different sets of weights, both of which would reproduce the original correlations to the same fairly close approximation. However, one or two instances have shown the weights from this method to correspond closely to those obtained by standard Doolittle procedures. Any disagreement is not likely to be of practical importance.

Index

☆ U. S. GOVERNMENT PRINTING OFFICE: 1947—703315