AFCAPS-FR-2019-0002

Technical Review and Analysis of the Army Air Force Aviation Psychology Program Research Reports

May 2019

Diane Damos, Ph.D.

Damos Aviation Services, Inc.
(DAS)

Prepared for:

Katie Gunther, Ph.D.

Air Force Personnel Center
Strategic Research and Assessment Branch

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 and by HQ AFPC Public Affairs and is releasable to the Defense Technical Information Center (DTIC).

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 JBSA-Randolph TX 78150-4747 or higher DoD authority. Please contact AFPC/DSYX Strategic Research and Assessment 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, Dr. Katie Gunther AFPC/DSYX, 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)***
05-02-2019	Final Report	

4. TITLE AND SUBTITLE Technical Review and Analysis of the Army Air Force Aviation Psychology Program Research Reports	**5a. CONTRACT NUMBER** HCaTs GS02Q17DCR0008
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER ITSS ID10180037
6. AUTHOR(S) Diane Damos, Ph.D.	**5d. PROJECT NUMBER**
	5e. TASK NUMBER 47QFAA18F0043
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) HQ AFPC/DSYX	**8. PERFORMING ORGANIZATION REPORT** AFCAPS-FR-2019-0002
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Personnel Center Strategic Research and Assessment Branch JBSA-Randolph AFB TX 78150	**10. SPONSOR/MONITOR'S ACRONYM(S)** HQ AFPC/DSYX
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFCAPS-FR-2019-0002

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

13. SUPPLEMENTARY NOTES

14. ABSTRACT

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), reviewed the series to identify tests and constructs that were not pursued after World War II, but have the potential to improve modern USAF aircrew selection and classification processes. Based on its review, DAS made eight recommendations. One involves examining the effect of including number incorrect, as well as number correct responses, on the factor structure of test batteries. Five recommendations suggest the development of new tests. One involves adding new questions to the physical science portion of the Air Force Officer Qualifying Test. Another pertains to identifying personality traits relating to carefulness and conscientiousness.

15. SUBJECT TERMS
Aviation Psychology, Testing, Aircrew Selection, Aircrew Classification

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**	33	210-565-5245

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

Table of Contents

Technical Review and Analysis

Introduction

Background

The Advanced Computer Learning Company (ACLC) contracted Damos Aviation Services, Inc. (DAS), to provide support for ACLC on Task 3 of the Strategic Personnel Research Program for the United States Air Force (USAF) *[Task Order # 47QFAA18F0043 under HCaTS Contract GS02Q17DCR0008, ITSS #: ID10180037].* As part of Task 3, DAS was to review the 19-volume Army Air Force's Aviation Psychology Program (AAP) Research Reports

> "*to identify systematic, prioritized recommendation of promising constructs/concepts worthy of further study for potential utility in Air Force selection and classification processes."*

This report will identify attributes or tests that for some reason were not pursued after World War II, but have the potential to contribute to modern aircrew selection and classification. Consequently, tests of attributes that are assessed in the current Air Force Officer Qualifying Test (AFOQT) are not discussed in this report unless a unique assessment approach occurred in the AAP reports.

The number, title, and general topic of each of the 19 volumes are given in Table 1. A quick review of this table shows that only *Reports No. 4, 5, and 7* are concerned directly with the development of selection tests. *Reports No. 8* through *13* deal with training issues for the various aircrew specialties, particularly with the criterion problem. Two others, *Reports No. 14* and *15*, deal with reassignment of aircrew members after combat tours. The other reports provide information on other tasks assigned to the AAP team, data collection and analysis challenges, and human factors issues in the design of aircraft displays and controls. Thus, only *Reports No. 4, 5,* and *7* seemed likely to provide tests or constructs worthy of further development for pilot selection. Nevertheless, DAS reviewed, page by page, the 7,500 pages contained in all 19 reports for tests and constructs that could potentially improve the current Air Force aircrew selection and classification system. This review identified additional tests and constructs in *Reports No.12* and *17* that will be discussed in the body of this report.

Table 1. Army Air Forces Aviation Psychology Program Research Reports by Number, Title, and Topic

Report Number	Title	Topic
1	Aviation Psychology Program in the Army Air Forces	Overview of program.
2	Classification Program	Organization of Aviation Psychology Program classification battery by month and year.
3	Research Problems and Techniques	Job analysis. Statistical approaches to determining the validity and reliability of tests and criterion measures.
4	Apparatus Tests	Description and results from all apparatus tests.
5	Printed Classification Tests	Description and results from all paper-and-pencil tests.
6	The AAF Qualifying Examination	Item development process for printed tests.
7	Motion Picture Testing and Research	Development of motion picture selection and assessment tests.
8	Psychological Research on Pilot Training	Development of criteria for pilot training.
9	Psychological Research on Bombardier Training	Proficiency assessment measures. Instructor selection.
10	Psychological Research on Navigator Training	Job analysis. Proficiency assessment measures. Instructor selection.
11	Psychological Research on Flexible Gunnery Training	Selection and training of gunners.
12	Psychological Research on Radar Observer Training	Job analysis. Proficiency assessment measures. Test validation.
13	Psychological Research on Flight Engineer Training	Job analysis. Selection test validity. Proficiency assessment measures.
14	Psychological Research on Problems of Redistribution	Issues of reassignment of personnel returning from combat.
15	Psychological Research Program in AAF Convalescent Hospitals	Evaluation processes and assessments for combat causalities.
16	Psychological Research on Operational Training in the Continental Air Forces	Analysis of aircrew duties and available criteria.
17	Psychological Research in the Theaters of War	Operational criteria development and specialized selection.
18	Records, Analysis, and Test Procedures	Formatting and analysis of test results.
19	Psychological Research on Equipment Design	Human factors issues in the design of controls and displays for aircrew.

Criterion Used in Test Development

To understand the validity data described in this report, the reader should be familiar with the process of aircrew classification. Men interested in qualifying as an aircrew member first took a general examination, the Army Air Forces Qualifying Examination (Flanagan, 1948, p. 53). Those who passed were designated as “aviation cadets” and were sent to a testing center where they received a flight physical. Applicants who passed the physical examination then completed the Aircrew Classification Battery and were assigned to pilot, navigator, or bombardier training.

Interpreting the data given in *Reports No. 4, 5*, and *7* can be problematic because different types of data were collected for the aircrew specialties at different points in the selection and training process. The data used for test development for any aircrew position often were obtained from aviation cadets. Descriptive statistics usually were based on aviation cadets. Validity data were obtained from the specialty in question (e.g., from pilots for pilot selection tests). The criterion used to determine predictive validity varied across specialties. For navigators, the criterion was pass/fail from advanced navigator training. For bombardiers, pass/fail from advanced bombardier training. Pilot training was divided into preflight, primary, basic, and advanced training (Thorndike, 1947, p. 32). Although the standard criterion for pilot selection tests in the classification battery was pass/fail from primary flight training, other criteria were used, such as pass/fail from advanced training.

Organization of This Report

The results of this review are organized by report. *Report No. 4* is the largest of the 19 volumes. Nevertheless, reviewing the information was straightforward, and the section needed no subdivision. The information found in *Reports No. 7, 12*, and *17* was limited and needed no subdivision. The amount of information obtained from *Printed Classification Tests Report No. 5* was extensive, and organizing this information in a meaningful manner for a modern reader was problematic. Remarks by Humphreys (1947, p. 45) indicated that each test presented in *Report No. 5* initially was categorized as a test of either intellect, perception, or temperament. The tests then were grouped by category into chapters, with at least three chapters for each of the three categories. Humphreys (1947) also noted that after work began on *Report No. 5*, the assignment of some tests was felt to be inappropriate, and the tests were moved to a different category and chapter.

Despite the reassignments, readers will find the structure of *Report No. 5* puzzling. Chapters that were assigned to one category appear to belong to another. Groups of tests within a chapter often seem to assess different attributes. Even tests within a group might assess different attributes, and the assignment of specific tests to chapters and groups often appears arbitrary. To aid understanding, DAS grouped chapters that assessed the same or similar attributes together, regardless of their category. Specifically, all of the chapters that are concerned with personality and motivation are reviewed together, as are all of the chapters dealing with spatial processing.

Apparatus Tests Report No. 4

Overview

The decision to examine the usefulness of apparatus tests for pilot candidate selection was made in October 1941 (Melton, 1947, p. 1). This decision was based on research conducted between World War I and World War II showing the usefulness of the Mashburn (Complex Coordination) Test (Mashburn, 1934), the Two-Hand Coordination Test, and the Rotary Pursuit Test for pilot selection. The decision was also based on the results of surveys conducted by the AAP team to identify the causes of failures in primary flight training. One of the major causes for failure pointed to poor “perceptual motor coordination” (Melton, 1947, p. 1). Consequently, the AAP team decided to examine a variety of apparatus tests, including the three mentioned above.

Although validation research on apparatus tests began in February 1942, the AAP team was surprised by an order issued in that month requiring them to select immediately the specific apparatus and printed tests to be used in the aircrew classification battery (Melton, 1947, p. 2). Because of their previously demonstrated utility in pilot selection, the Complex Coordination Test (CM701B) and the Two-Hand Coordination Test (CM101B) were selected for the first battery, which was fielded in August 1942. Wartime shortages of mechanical parts limited the options for other tests. A reaction-time test, a fine-motor-coordination (finger dexterity) test, and a test of hand steadiness also were included in the first battery because the parts needed to build their apparatus could be obtained easily. Although the team knew that the Rotary Pursuit Test (CM803B) measured a unique skill, it was added to the battery only after the team ascertained that parts were available to construct its apparatus. Thus, its inclusion in the battery was delayed until December 1942.

As noted by Melton (1947, p. 4), the number and types of apparatus tests in the classification battery remained surprisingly stable throughout the war. Revisions occurred periodically, and occasionally a test in the battery was replaced with a new test that assessed the same construct. The only major addition to the apparatus battery occurred in June 1945, when the Pedestal Sight Manipulation Test (CM824A) was added for the express purpose of predicting B-29 gunnery training performance. The stability of the apparatus battery might be attributed to the success of the apparatus tests. Throughout the war, the apparatus tests were weighted almost as heavily as the printed tests in the prediction equations.

One goal for the AAP team was to break down complex task performance into its components and measure the performance on each component separately (Melton, 1947, p. 985). The hypothesis was that fine-grained measures of performance could reveal attributes that contribute to overall performance but are not reflected in gross (total) measures of performance. This goal was not achieved because of calibration problems with the testing units and the limited performance measures that typically could be obtained. The common measures included reaction time, number of items completed, and time on target. All of these assess only gross performance. The AAP team attempted to obtain more fine-grained measures of performance that could reveal new attributes. One such attempt concerned the Two-Hand Coordination Test (CM101B). The AAP team attached an impact recorder to the device to measure smoothness of control. The smoothness variable changed little with practice on the device and had high inter-trial correlations. Candidates who made smoother movements had a slight tendency to achieve better scores on other apparatus tests, a

finding consistent with a distinct attribute (Melton, 1947, p. 237). However, Melton noted constant calibration problems with the Two-Hand Coordination Test, with different test units producing very different mean scores. He believed that the calibration issue was at least partly the cause of a non-significant correlation between pass/fail from flight training and the smoothness measure.

The AAP team also attempted to obtain a fine-grained measure of performance on the Rudder Control Test (CM120B). For this test, the team obtained a measure of the "amount of movement" used to perform the test. Like the smoothness-of-control measure, the amount-of-movement variable had high inter-trial correlations. Two preliminary studies showed that the movement variable predicted pass/fail from primary flight training. However, this result was not replicated in a larger study. Melton (pp. 454–455) again found large performance differences among copies of the testing apparatus used in the larger study and recommended more research on the amount-of-movement variable.

Assessment of Apparatus Tests

There is no question that apparatus tests were successful in predicting the outcome of aircrew training. Nevertheless, calibrating and maintaining the apparatus was a constant problem that had deleterious effects on the predictive validity of some of the tests (Melton, 1947, p. 981). The AAP team's attempt to decompose complex performance into its components and study performance on the components achieved limited success because of the calibration problems. Melton (1947, p. 985) acknowledged the shortcomings of the apparatus test research and development program. He advocated developing a selection test that would be complex, require division of attention, and need fine-motor adjustments for good performance. He also advocated more research on individual differences in psychomotor performance.

The continued importance of psychomotor tests for pilot selection is evident in the fact that computerized compensatory and pursuit tracking tasks and their combination are included in the U.S. Air Force's Test of Basic Aviation Skills (TBAS) battery (Carretta, 2005; Ree, 2003). These are the modern equivalent of the Rotary Pursuit Test (CM803B) and the Rotary Pursuit Test with Divided Attention (CP410B). The calibration issues that plagued the World War II effort effectively have been eliminated by computerization of the tests. Moving control devices, such as control sticks, still need maintenance, but computerization has minimized the maintenance problems. Computerization also allows a large number of variables to be recorded at sampling frequencies that were impossible during World War II. The development of new analysis techniques, such as fast Fourier analysis, produces performance measures unknown in World War II.

Because "apparatus" tests that assess eye-hand and eye-hand-foot coordination are currently in the pilot selection battery, there is no need to add a comparable test. The Air Force needs to add a much more difficult test that stresses the examinee's time-sharing skills as well as their psychomotor performance. It needs to be cognitively challenging and have high face validity. The best test for this is a computerized version of the Multiple Control Stress Test.

The following description is the best interpretation DAS personnel could make about how the test functioned.

This test had two, two-dimensional tracking tasks, each with its own display. The larger display had two spots and would be classified today as a pursuit-tracking task. One spot, the target, was moved

horizontally by the testing mechanism. It was moved vertically both by the testing mechanism and by the control stick, which was held in the candidate's right hand. How these two inputs were combined to produce vertical movement is unknown. The use of control stick movement as an input to the target might be thought of as a type of cross-coupling. The other spot, the cursor, was completely controlled by the stick. The control/display relation of the stick was reversed from that of an aircraft (i.e., pushing the stick forward caused the cursor to move up). The candidate's task was to keep both dots centered vertically on the display. The reader should notice that this is different from most pursuit tracking tasks, which require the candidate to keep the cursor on the target regardless of the target's location on the display.

The second tracking task had one spot, which implies that it was a compensatory tracking task. The testing mechanism moved the spot only in the vertical direction. A mock throttle manipulated by the candidate's left hand was used to compensate for the vertical movement of the spot. Input for the horizontal movement of the spot came from the stick in the candidate's right hand. Again, this implies a cross-coupling, but this time between the control stick input and the horizontal movement of the spot of the second tracking task. No information is provided about exactly how movement of the control stick affected the horizontal movement of the spot of the second tracking task. The horizontal movement of the spot was compensated for by rudders, which had the same control/display relation as an aircraft.

The third task required the candidate to monitor four pairs of lights, which apparently could be either red or green. Some type of response device was placed in front of the candidate with switches for the lights. When a light turned red, the candidate had to move a switch corresponding to the light. If the candidate failed to respond to the light change within 15 s, a loud, unpleasant noise sounded in their headset.

The candidate began with two trials on only the first tracking task. Then the candidate performed the second tracking task concurrently with the first task for two trials. On the fifth trial, a lag of undisclosed length was introduced into the forward/backward movement of the control stick. The third task was added during Trial 6. The next two trials were easier, with Trial 7 requiring only dual-task tracking without the lag. Trial 8 was the same as Trial 1. Only limited testing was conducted on the Multiple Control Stress Test, and no descriptive data were provided.

Printed Classification Tests Report No. 5

Overview

The printed classification tests were divided into three major categories: intellect and information, perception, and temperament (personality) (Humphreys, 1947, p. 45). Tests of intellect included assessments of verbal, math, reasoning, mechanics, judgment, foresight and planning, integration, memory, visualization, and information. Tests of the first four areas will be included in the first section below. Judgment and foresight and planning tests are discussed in the second section, followed by sections on integration and memory. Attention, which is discussed next, was considered by Humphreys to be related to perception. However, for the purposes of this report, it will be discussed separately. The section on perceptual tests is followed by the section on spatial tests.

Tests of visualization were included with other tests of intellect in *Report No. 5*, but for the purposes of this report, they are included in the spatial section. “Information” refers to biographical data. Information tests are included in the final section, which deals with personality.

Verbal Skills, Mathematical Skills, Reasoning, Mechanics

Overview. Tests of vocabulary and reading comprehension were included in the classification battery to ensure that the candidates could comprehend and remember the type of material presented in ground school. A commercial vocabulary test and a reading comprehension test developed in house (CI614G) were included in the December 1942 battery. The reading comprehension test was revised several times but remained in the aircrew classification battery throughout the war. The commercial vocabulary test was replaced in August 1942 by the Technical Vocabulary Test (CE505C).

Tests of mathematical skills were introduced in the first objective examination administered in 1941 (Davis, 1947a). Some of these tests assessed knowledge of arithmetic, algebra, and trigonometry. Others were classified as “computation tests,” which assessed only arithmetic skills. Both types of tests were in the classification battery until December 1942, when the computation test was replaced with a numerical reasoning test, Numerical Operations (CI701A).

Most of the reasoning tests were designed for pilot selection. Because the mathematical skills tests had poor predictive validity for pilots, the AAP team developed several nonverbal, nonmathematical reasoning tests. These tests are similar to the Raven’s Progressive Matrices (Raven, 1938), and several were modifications of commercially available tests.

Tests of mechanics consisted of tests of mechanical comprehension, mechanical information (e.g., “What device drives a fuel pump?”), and basic physics. The Mechanical Principles test (CI903A), a test of mechanical comprehension, had high predictive validity (r_{bis}[1] > 0.33) for pilots and to a lesser degree for navigators. Mechanical Principles entered the aircrew classification battery in December 1942 and remained in the battery throughout the war. Mechanical Information (CI905B) also was predictive for pilots. This test entered the battery in September 1944 and remained in the aircrew classification battery throughout the war. The only physics test that was validated assessed mechanical experience rather than knowledge of physics. No suitable physics test was developed before the end of the war.

Assessment of verbal, mathematical, reasoning, mechanics tests. Many of the verbal, mathematical, and mechanical tests were derived from commercial sources (Humphreys, 1947, p. 49). The verbal and mathematical tests described above have derivatives in the current AFOQT. None of the tests is noteworthy for unusual content.

Mechanical comprehension and information tests were predictive for pilots in World War II. Developing a good test of mechanical comprehension and/or information was difficult because it was, and is, affected heavily by prior experience (Carroll, 1993, p. 526). Such tests are no longer in the AFOQT. The physics tests developed in World War II were never included in the classification battery. Given the highly technical nature of modern aircraft and operations, it may be worthwhile to re-examine the usefulness of these tests.

[1] r_{bis} is the uncorrected biserial correlation.

Judgment and Foresight and Planning

Overview. The original pilot job analyses conducted by the AAP team included extensive analyses of review board records and interviews with flight instructors about the causes of failures in pilot training (Walton, 1947, p. 2). One commonly cited reason was "poor judgment" (Fruchter, 1947a, p. 123). This chapter described attempts to develop a practical judgment test for pilot selection. Early in the process, the AAP team recognized that scores on the judgment tests were affected by the candidate's level of mechanical knowledge. Subsequent efforts attempted to eliminate any contribution of mechanical knowledge to test performance (Fruchter, 1947a, p. 127). Word fluency also was examined for its potential contribution to judgment questions requiring specific facts and experiences. One judgment test, Practical Judgment (CI301C), was included in the aircrew classification battery in September 1944 and remained in the battery throughout the war.

Like judgment, lack of foresight and planning was identified as a major contributor to failure in primary flight training (Mock & Guilford, 1947, p. 157). Most of the tests in this chapter dealt with route planning, such as through a town, or determining the most efficient path to perform a skywriting assignment. A mock game of "dots and boxes" also was included in this chapter, with the candidate anticipating both contestants' moves. None of these tests was included in the classification battery.

Assessment of judgment and foresight and planning tests. Although a judgment test was included in the selection battery in 1944, the exact attribute assessed by this test is unclear. The judgment tests developed by the AAP team were affected heavily by the candidate's mechanical experience and other prior experiences. These results suggest possible ethnic and gender differences if the same development process were pursued today. Situational judgment[1] tests appear to be a better option.

Factor analyses of the foresight and planning tests were inconclusive. Two previously unknown factors were found but were not easily identified. Several of these tests had high face validity, and some had game-like properties. Nevertheless, any new development would require a major research effort to understand foresight and planning.

Integration

Overview. "Information integration" was understood differently in World War II compared with what it means today. Currently, the term usually implies the processing of continuous information from multiple sources that need to be combined to make a response. During the war, the idea of continuous information processing appears to be lacking.

The integration tests supposedly required the candidate to divide his attention among several information sources while retaining a large amount of information in short-term memory. The goal was to test a candidate's ability to integrate the information quickly and accurately. This goal was never achieved because six of the seven tests were paper-and-pencil tests. Thus, the only source of information was the instructions, and the only way to control the presentation of information was to

[1] A Situational Judgment subtest was added to AFOQT Form T.

have information appear at different points in the instructions or in the test questions themselves. The seventh test used motion pictures to present three successive groups of three colored squares. The candidate had to remember how many squares of each color were presented in the trial.

Five tests had predictive validity data for pilots. Validities ranged from approximately 0.0 to corrected biserial correlations of approximately 0.24. Only the motion picture test had validity data for navigators, which showed r_{cbis} [2] = 0.30 for training completion.

Assessment of integration tests. The concept of these tests was good despite the team's very limited ability to control the presentation of information. Clearly, this type of test could be computerized, which would allow multiple sources of information to be presented at different points in time. A high short-term memory load could be induced by extensive instructions that needed to be retained to execute the tasks. The level of realism could be easily varied.

Memory

Overview. Memory issues were prominent in review board records, accounting for about 24% of the reasons for failure in primary flight training (Shirley, 1947, p. 227). Even in advanced single-engine pilot training, about 52% of the failures included some form of memory deficit. Memory also was considered an important attribute for navigators and bombardiers; it received above-average ratings in required attributes.

The understanding of memory in the early 1940s was substantially different from what it is today. Earlier academic research had resulted in the development of various ways of testing memory, such as the paired associate and memory span paradigms, and factor analyses of data from these paradigms demonstrated the existence of more than one memory factor (Shirley, 1947). The team appeared to recognize short-term and long-term memory systems but was severely limited in the types of memory tests that could be included in the battery by the requirement for group testing and machine scoring.

The majority of tests were "pictorial memory" tests that required the candidate to study maps and answer questions about the location of landmarks. A few of these tests involved recognition of aircraft or ship silhouettes. Another type of test was the "symbolic memory" test. Only three of these tests were constructed. One required the candidate to remember detailed information about a tactical plan. The second was similar, except the information to be remembered described a geographical area. The stimuli for the third test were three-character designations of aircraft. The candidate was given the designation and had to recall the name of the aircraft. No factor analyses were conducted on any of these three tests, and validation data were very limited.

To obtain a better understanding of memory, the AAP team constructed two batteries consisting of seven tests from the aircrew classification battery and either five or six memory tests developed prior to the summer of 1942. The sample sizes were small for both batteries: 179 bombardiers and 239 aviation cadets, respectively. The team appears to have misinterpreted the results of analyses conducted on both data sets. They apparently thought they found a paired associate memory factor and a visual memory factor. In reality, the factors represent method (paradigm) variance,

[2] r_{cbis} is the biserial correlation corrected for range restriction.

not different memory systems. The team was confused further by the failure of at least one of the memory tests to load on what they thought was a general memory factor.

Assessment of memory tests. Many of the memory tests were dismissed from consideration because of administration issues, low reliability, or low validity. Based on their interpretation of the factor analyses described above, the AAP team cautiously suggested that some memory tests should be added to the aircrew classification battery to increase its predictive validity. However, no memory test was ever included in the battery.

Two of the symbolic memory tests are worth consideration. The first is the Geographical Memory test (CI508AX). At the start of the test, the candidate was given a one-paragraph description of an area that contained railroads, factories, highways, etc. The candidate had 7 minutes to study the information and then was asked a series of questions about the direction and distance between certain important features. This test forced the candidate to construct a detailed mental map and retain the map for 10 to 12 minutes.

In the second test, Memory for Tactical Plans (CI509BX), the candidate listened to a briefing of a bombing mission, read aloud by the test administrator (Shirley, 1947, pp. 255–257). Examinees were told that they would be asked about certain details of the mission in a few hours. Between 2 and 3 hours later, examinees answered multiple-choice questions on details of the briefing. This test was different from the other memory tests in that the information to be remembered was presented orally and the retention period was several hours. The validity for primary flight training completion (pass/fail) was somewhat low, $r_{cbis} = 0.19$. However, the reliability was 0.68 (alternate forms method, corrected for length) and the communality was only 0.47. Shirley (1947, p. 257) felt that this test would be a valuable addition to the aircrew classification battery. However, the oral administration of the material made it impractical.

Set and Attention

Overview. Job analyses conducted by the AAP team identified sustained attention, attention under distraction (divided attention), and change of set as three important abilities for success as a pilot and, to a lesser degree, for navigators and bombardiers (Fruchter, 1947b, p. 541). Research on these topics was hampered by a limited understanding of attention in the 1940s and by the need to assess attention using paper-and-pencil tests administered in a group-testing situation. Fruchter (p. 563) noted that there was no concerted effort in this area, and few studies were conducted on set and attention despite their apparent importance. Only two sustained attention tests were developed. The first required a large amount of information to be kept in auditory short-term memory. The other, Following Directions (CP402A), consisted of instructions on how to fill out an answer sheet. The instructions were written in between the blanks on the answer sheet and changed as the candidate worked down the answer sheet. Both tests had low predictive validity for pilots ($r_{cbis} < 0.15$). Data for navigators were collected only on the Following Directions test and showed good predictive validity ($0.24 < r_{bis} < 0.43$).

The AAP team created four “attention under distraction” tests. The first three required the examinee to perform a primary task in the presence of distracting items. Examinees were scored on the primary task and on their recall of the distractors. The first test also asked examinees to rate their confidence in his answer. No objective measures for the first test were predictive for either pilots

or navigators, and only the confidence rating was predictive for navigators (r_{cbis} < 0.29). No validation data were collected on the second test. The third test had validation data only for pilots and showed poor predictive validity (r_{cbis} < 0.11).

The fourth test required the examinee to trace a maze with their left hand while tracing a different maze with their right hand. This test had good predictive validity for primary flight training completion (r_{bis} = 0.29). No validity data were collected for other aircrew positions. This unusual test was constructed in an attempt to develop a paper-and-pencil test that assessed the same attribute as an apparatus test, which in this case was the Two-Hand Coordination Test (CM101B). No data were presented to support this notion.

A medical and psychological unit developed the change of set tests, not by the AAP team. All eight tests were designed to measure flexibility of attention. All of the tests involved an initial set of problems that required a specific strategy to solve. The second set of problems could be solved using the same strategy as the first set or a simpler strategy. The third set could be solved using either of the two preceding strategies or by an even simpler strategy. The dependent measure was the number of problems solved before adopting the simpler strategy. The eight tests had low intercorrelations, raising questions about their validity. Predictive validity data were collected only for pilots and on only one of the eight tests. These data showed a moderate predictive validity (r_{cbis} < 0.21). Today, these tests can be recognized as tests of perseveration.

Assessment of set and attention tests. The limited validity data collected from the sustained-attention and divided-attention tests on pilot trainees showed low predictive validity (r_{bis} < 0.15), although the validity for navigator trainees was somewhat higher (r_{bis} < 0.24). Of the attention tests, Fruchter (1947b, pp. 563–564) considered only the Following Directions test as promising, and it was predictive only for navigators. However, he believed that this test assessed an integration attribute, not sustained attention. The change of set tests assessed perseveration, which today is not considered important for a pilot. Additionally, commercial tests are available to measure perseveration if the Air Force decides to pursue this attribute.

Perceptual Tests

Overview. The perceptual category includes chapters on perceptual speed, form perception, and size and distance estimation. The chapter on perceptual speed includes two types of tests. The first type hypothetically involves speed of comparison and includes the Speed of Identification test (CP610A). This test was included in the aircrew classification battery beginning in March 1942. The only interesting aspect of this test is that it might have included a mental rotation element that was not identified by the AAP team. The second type of perceptual speed test assessed clerical speed. This category included the Table Reading test, which is still in use today. All of these tests required graph, table, or dial reading and were very similar. These tests showed moderate to high validity for navigators ($0.20 < r_{bis} < 0.50$) and low to moderate validity for pilots ($0.00 < r_{bis} < 0.25$).

The form perception chapter includes tests of illusions and what are now known as "hidden figures" tests. Other tests required examinees to assemble parts of a drawing or photograph to make a meaningful picture. The illusions tests attempted to measure individual differences in the experience of visual illusions. The illusion tests suffered from low reliability, and no validity data were obtained on any of these tests. The assembly and hidden figures tests generally had low predictive validities for pilots, with r_{bis} frequently less than 0.15.

The size and distance estimation chapter contains angle and proportion estimation tests as well as distance estimation tests but no tests of size estimation. Three of the distance estimation tests—Shorter Line (CP606), Nearest Point (CP607), and Shortest Path (CP608)—were copyrighted by a commercial group and included in the February 1942 classification battery.[3] All three had moderate predictive validities for pilots ($r_{bis} < 0.20+$) but suffered from low reliabilities. All three were removed from the battery by April 1942.

Three angle estimation tests were developed. The first, which had a moderate predictive validity for pilots ($r_{bis} = 0.20$), required examinees to estimate the angle between two lines drawn on paper. The second test proved too difficult to use. No data were obtained on the third, an angular judgment test. The only proportion estimation test included in this chapter suffered from low reliability and poor predictive validity.

Assessment of perceptual tests. The perceptual speed tests are unremarkable in both in their content and their development process. One of the clerical speed tests, Table Reading, is still in the AFOQT, and there is little reason to consider another such test. The form perception tests generally demonstrated low predictive validity. See Carretta (1987) and Olea and Ree (1994) for more recent data on a hidden figures test. The angle and proportion estimation tests show no promise for continued development.

The chapter authors, Lacey and Shirley (1947, p. 476), suggested continued experimentation with distance estimation tests. Today, the usefulness of such tests is questionable given the distance and altitude measurement capabilities of modern avionics.

[3] Lacey and Shirley (1947, p. 448) incorrectly stated that these three tests were included in the April 1942 classification battery. These tests were removed from the battery in April 1942.

Perhaps the most important finding from these chapters concerns the use of scoring formulas. Lacey and Shirley (1947) demonstrated that the correlations between one of the perceptual tests and other tests in the battery differed substantially when number correct was used as the dependent variable versus number correct adjusted for errors (p. 460). To them, the changing correlations raised issues about the factorial composition of the test (p. 475). Additionally, they noted an unusual positive correlation between number correct and number wrong for the Estimation of Length test (CP631A), again highlighting scoring issues associated with the use of the number correct versus an adjusted number correct. This issue will be discussed further in the section on personality and motivation.

Spatial Tests

Overview. In their introduction to spatial tests, Howe and Zimmerman (1947) noted that "spatial ability" was a relatively new term in the early 1940s. Thurstone first identified spatial ability as a primary mental ability in 1938, only a few years before the outbreak of World War II (Thurstone, 1938). The AAP investigators involved with developing the printed selection tests were aware of Thurstone's work on both primary mental abilities and factor analysis (Thurstone, 1935). Consequently, these investigators assumed the existence of one spatial ability, and this assumption underpinned test development early in the war (Howe & Zimmerman, 1947, p. 477).

Howe and Zimmerman (1947, p. 477) implied that the development of spatial tests was not an initial priority of the classification program. Research on this topic began only after factor analyses conducted on printed tests of planning and the Complex Coordination test revealed an unanticipated factor. This factor was not well defined, but was tentatively identified as a spatial ability. The existence of this factor clearly was a surprise because neither the planning tests nor the Complex Coordination test was assumed to require a spatial ability. The identification of this new factor seems to have spurred research into spatial ability. Howe and Zimmerman (1947, p. 477) discussed the issues involved with identifying this ability and subsequent efforts to develop tests that would load on this new factor. These efforts apparently were successful because later factor analyses clearly showed two spatial factors.

Carroll (1993, p. 308) remarked on the confusion that exists in identifying factors in the visual perception domain, a domain that includes what are now regarded as spatial factors. This same confusion is evident in the categorization of tests in three chapters in *Report No. 5.* All three chapters deal with spatial abilities. Nevertheless, the visualization chapter was included in the intellect category, whereas the spatial and orientation chapters were in the perceptual category. Additionally, the assignment of a specific test to a chapter often appears inappropriate to a modern reader. The spatial chapter is particularly problematic because several of the tests included in this chapter appear to involve no spatial abilities. Because the assignment of tests to chapters and chapters to categories appears arbitrary, only one assessment will be given for all three chapters.

Spatial. The spatial ability tests were divided into two subgroups: directional discrimination tests and positional discrimination tests (Howe & Zimmerman, 1947). Instrument Comprehension (CI615B–CI616C), a directional discrimination test, was included in the classification battery beginning in November 1943 and remained in the battery until at least June 1945. Of the remaining five directional discrimination tests, only one had validity data. That test, Aerial Orientation (CP520A), required examinees to compare a view of the cockpit with five photographs of an aircraft

in different orientations and choose the one matching the cockpit display. It was suggested as a substitute for Instrument Comprehension because it had high predictive validity for both pilots and navigators (r_{bis} = 0.29 and 0.31, respectively).[4] However, there is no indication that Aerial Orientation was ever used in the classification battery.

The positional discrimination tests were studied in an attempt to identify a second spatial factor. One was a mirror versus standard orientation test, the second was a mental rotation test, and the third required the identification of left versus right hands. All three tests seemed to be undergoing refinement and validation at the time the research was discontinued. The validity data are difficult to interpret because several of the tests were divided into parts, and the validity was reported by part and by scoring formula (number correct, number wrong, number correct minus number wrong). Howe and Zimmerman (1947) felt that the work on all three tests was inconclusive, and none was recommended for use as selection instruments.

Orientation. Surveys examining the causes of failure in flight training (Youtz, 1947, p. 41) identified low orientation ability as a cause for early elimination from flight training (Lacey & Niehaus, 1947, p. 511). This identification seems to have resulted in the rapid development of orientation tests; the first orientation test, Spatial Orientation (CP501A), was included in the aircrew classification battery by December 1942 and remained in the battery until at least June 1945.

Lacey and Niehaus (1947, p. 511) defined orientation abilities as "the abilities to determine one's bearings with respect to points of the compass and to maintain appreciation of one's location relative to landmarks in the environment." Two types of orientation tests were developed in keeping with the definition given above. The first type dealt essentially with compass orientation; the second, with orientation relative to the ground. The second type was referred to as "pattern orientation."

Several of the compass orientation tests showed useful validities for pilots. One of these, Directional Orientation (CP515), had six versions (A–F). Versions D, E, and F used photographs rather than drawings to allow more flexibility in stimulus presentation. These versions were developed relatively late, and no validity data were reported. Versions B and C had predictive validities exceeding 0.35 for both pilots and navigators after correction for dichotomization of the training completion criterion. Version A had corrected predictive validities exceeding 0.21 for pilots.

Another compass orientation test, Following Oral Directions (C1651AX), had a unique presentation method. This test had four levels of difficulty. For all four levels, examinees assumed that they were flying an aircraft. At the start of each test item, they were told the cardinal direction they were flying. After a short interval, they were told that they were being attacked from the left, right, front, or rear. Attacks from the left or right required a 90° turn in the opposite direction. An attack from the front required a 180° turn. If the attack was from the rear, examinees were not to alter course. In the simplest version of the test, examinees were attacked by only one aircraft and maintained altitude. In the most difficult version, the examinee was attacked by two aircraft successively and had to change both heading and altitude.

[4] The validity data cited here are for pilots in primary training and unidentified navigators. This test is also known as Orientation to Landmarks. Additional validity data are discussed in sections of *Report No. 12* and *Report No. 17.*

The interval between being given the initial heading and the onset of the attack(s) varied with the difficulty level of the test item. At the end of the trial, the examinee reported their current heading. The test description does not indicate if altitude changes were also reported. All instructions were oral and given by phonograph.

The validity data for Following Oral Directions was collected between mid- and late 1944. The predictive validity was moderate for pilots, with $r_{cbis} > 0.24$. Scores on this test were included in a factor analysis using scores from the existing aircrew classification battery. Only 8% of the variance of Following Oral Directions was accounted for by the spatial relations factor and only 3% by the verbal factor. A second version was developed (CP651BX), which was longer and faster paced. The predictive validity for pilots was about the same, $0.22 < r_{cbis} < 0.29$, and slightly higher for navigators. Few other statistical data were given. However, a third version was developed, but no data were available at the time *Report No. 5* was printed.

The pattern orientation tests apparently were more difficult to develop, and tests often went through multiple iterations. Few validity data were collected for pilots, and several were still under development when research stopped. The exception was Spatial Orientation, which was discussed earlier. However, as noted by the authors, its usefulness lay in its assessment of perceptual speed, not in its assessment of an orientation ability (Lacey & Niehaus, 1947, p. 539).

Visualization. The same surveys of flight failures that identified orientation as an important ability also identified visualization as important (Youtz, 1947). In these surveys, "visualization" was concerned with controlling the flight path of the aircraft by using ground reference points. The visualization tests developed by the AAP team were designed to assess basic questions such as whether visualizing two-dimensional versus three-dimensional objects required different abilities (Zimmerman, 1947). Two groups of tests were developed, visual manipulation and visual completion. The visual manipulation tests involved movement, such as mentally rotating an object or folding a flat surface. The completion tests required either the "ability to visualize the completion of a design or the extrapolation of a line or a path" (Zimmerman, 1947, p. 292).

Six visual manipulation tests were constructed. Predictive validity data for pilots were available on four of the tests. The validity for pilots was low ($0.10 < r_{cbis} < 0.20$). Two of the tests also had data on navigators, which showed high validity ($r_{cbis} > 0.30$). Only one visual completion test was developed, Flight Path. In this test, the candidate was shown a drawing with several aircraft, each of which was flying a circular path. Only a small part of each aircraft's path was depicted. The drawing also contained numbered points. The candidate had to determine which point each aircraft would fly through if it continued on its current path. The predictive validity of Flight Path also was low ($r_{cbis} < 0.15$).

Assessment of spatial, orientation, and visualization tests. Only one test described in these three chapters, Spatial Orientation, was included in the aircrew classification battery, and no other tests were recommended. Only Lacey and Niehaus (1947) felt that more test development and validation might be useful. This lack of support for test-specific developmentmay be attributed at least partly to the limited understanding of spatial abilities at the start of the war. This limited understanding had two consequences. First, the research program lacked a theoretical basis to guide test development. Second, some of the tests measured factors other than the ones they were designed to measure. For example, both Spatial Orientation I and

II (CP501A–CP503B), which were used in the aircrew classification battery for more than three years, measured predominantly perceptual speed, not spatial orientation (Lacey & Niehaus, 1947, p. 529).

The “Spatial” chapter includes the Instrument Comprehension test, which is currently in the AFOQT. This chapter also contains tests assessing mental rotation and left-right discrimination. Both attributes have been extensively studied over the last 30 years. Consequently, none of the tests in this chapter shows promise for renewed development.

In the “Orientation” chapter, only the compass orientation tests were of interest, several of these tests had moderate to high corrected predictive validities for pilots and high corrected validity for navigators. The problem with these tests is the attribute being assessed. Most of the compass orientation tests required the candidate to determine a direction using a magnetic (whiskey) compass that did not have north in the nominal position. Modern avionics systems can show either track up or north up, and even legacy avionics systems contain a directional indicator that minimizes the need for mental rotations and calculations to determine directions. Additionally, all of the compass orientation tests appear to have required mental rotation, which can be assessed more effectively using other tests.

The exception to the discussion above is Following Oral Directions (C1651AX), which is very different from the other compass orientation tests. The need to build a visual representation completely from verbal information is unique and worth considering for a new selection battery.

One visualization test, Flight Path, is promising despite its low predictive validity. Zimmerman (1947) correlated performance on this test with performance on other unspecified “existing” tests (presumably tests in the classification battery on or after July 1944) and found that it correlated poorly with all the other tests. He assumed, therefore, that it would contribute unique variance to the battery. Because of its high face validity and its poor correlation with other tests, Flight Path is worth considering for a new battery.

As noted earlier, only one test in these three chapters—“Visualization,” “Spatial,” and “Orientation”—had sufficient validities to be included in the aircrew classification battery. Although little was understood about spatial abilities during World War II, other problems might account for the general lack of predictive validity. Several of these tests, such as Flight Path and Stick and Rudder Orientation, attempted to represent motion in a static environment. Many others had to depict a three-dimensional world on a two-dimensional surface. Such problems can be easily circumvented today with existing technology.

Personality and Motivation

Overview. Six chapters in *Report No. 5* are devoted to testing personality and motivation. The chapter titles are confusing because three reflect the method of assessing personality—personality inventories, clinical methods, and biographical inventories—discussed in the chapter.

Two others refer to the attribute assessed. The sixth, "Information Tests," is concerned with the examinee's family history and with their knowledge of day-to-day activities, such as how to drive a car.

Between the wars, the Army conducted research using personality inventories. Because the validities were low, the Army was unwilling to commit significant resources at the beginning of World War II to the development of new personality inventories. Thus, the validation effort described in *Report No. 5* was limited to 12 commercially available personality inventories and 3 that were developed in house. None of the tests was found to be sufficiently predictive to warrant immediate use, and only one, the Guilford-Martin Personnel Inventory, was recommended for continued development (Cerf, 1947). There is no evidence that any further development occurred.

The clinical assessment methods included projective methods, one-on-one interviews, and "stress tests." Stress tests attempted to increase the candidate's stress level through negative comments made by the test administrator and limited response times. All of the clinical methods tests were labor intensive and demonstrated poor predictive validity for pass/fail from flight training or combat effectiveness.

Work on a biographical inventory began before the start of World War II and was considered promising by the start of the war. Development of a biographical inventory continued throughout the war, with the AAP team producing several versions of the inventory. Scoring keys were developed empirically. Version CE602D entered the classification battery in July 1943 and remained in the battery until at least June 1945. Separate scoring keys were used for pilots and navigators.

The first chapter concerned with attributes looked at five components of personality. Few validity data were available for tests of four of these attributes: masculinity, fear and tension, confidence, and social intelligence and leadership. Only tests of confidence were thought to be worth further development. The fifth component, carefulness, was believed to be important for navigators and was assessed by three tests. A factor analysis was conducted using both number correct and number wrong for each of the three tests of carefulness and the results of 11 tests from the classification battery (Davis, 1947b, pp. 686–695). The 11 tests included 6 apparatus tests and 5 paper-and-pencil tests. For the five paper-and-pencil tests from the battery, only number correct was included in the factor analysis. Data from 354 unclassified aviation candidates were analyzed. A total of 12 identifiable factors were extracted. Factor 1 was a numerical factor and was defined by the number correct from the three carefulness tests and two quantitative skills tests. Factor 3 had loadings only from the number wrong on the three carefulness tests. Factor 8 was defined by two apparatus tests and the mechanical ability tests as well as by the number correct on the three carefulness tests. Factor 10 was defined by four of the apparatus tests and the number correct on the three carefulness tests. Davis believed that Factor 3 was a carefulness factor (p. 691).

Although no validation data had been obtained for this attribute when *Report No. 5* was published, Davis (1947b, pp. 694–695) felt that a validation study was warranted.

The second chapter dealing with attributes was concerned with motivation. Many of the motivation tests were aircrew assignment preference inventories. Others assessed attitudes, particularly about the conduct of the war. Most of these tests had little relation to training outcome except for navigators, where the preference inventories were predictive.

Tests in the sixth chapter, "Information Tests," either collected biodata to assess knowledge, skills, and motivation or were tests of general knowledge. This chapter was included in the "Intellect" category because several tests assessed the examinee's technical knowledge, such as how to operate a motorcycle throttle, and his technical vocabulary. As noted by Humphreys (1947, p. 48), the assignment of the Information Tests to the intellect category was "somewhat arbitrary." Two information tests were included in the aircrew classification battery at some point. The first, a technical vocabulary test, was introduced in August 1942 but removed before July 1943. The second, the General Knowledge Test (CE505D), was introduced into the classification battery in July 1943. It was revised several times and remained in the battery throughout the war.

Assessment of personality and motivation tests. There is little new to be gained from the work on personality and motivation. Two of the personality assessment methods (inventories and clinical methods) were singularly unproductive because they focused on attributes of personality that affected performance in combat, such as fear, emotional control, self-confidence, leadership, and motivation to engage in combat. Developing instruments to assess these attributes requires good criteria. However, as Guilford (1947, pp. 570–571) observed, the developers had only very limited combat criteria, and the quality was "unsatisfactory at best" (p. 570). Relevant training data also were limited, with only 1.5% of training failures attributed to fear of flying. Additionally, developing any selection instrument requires an adequate sample. Guilford noted a high attrition rate from the time of testing to the completion of air combat duty. This attrition, combined with the poor quality of records in the combat areas, effectively doomed any validation attempt for these two assessment methods.

The third personality assessment method resulted in the biodata test described in the "Biographical Data" chapter. It was predictive, but both the development process and the content were unremarkable. The biodata test can be seen as related to the technical vocabulary and general knowledge tests developed in the information test chapter. These tests were predictive, and derivatives remain in the AFOQT today.

Two chapters dealt with personality attributes. One examined motivation through assignment preference inventories. These inventories had moderate success in predicting training outcome for navigators but, again, are unremarkable.

The other chapter was concerned with five personality attributes. Four of these attributes were not considered for continued development. The fifth, carefulness, warrants further development. As noted earlier, a factor analysis was conducted using both number correct and number wrong for each of the three tests of carefulness and the results of 11 tests from the classification battery (Davis, 1947b, pp. 686–695). Factor 3 had loadings only from number wrong on the three carefulness tests.

Davis (1947b, p. 694) noted the significance of having one factor (Factor 3) defined entirely by

error scores. He also noted that no factor had significant loadings from both the number correct and the number wrong from any of the three tests. These results suggest that number correct and number wrong may assess different attributes. Support for this notion is provided by Mount, Oh, and Burns (2008), who demonstrated that the number wrong on a perceptual speed test was related to citizenship behavior, whereas the number correct reflected task performance. Davis suggested analyzing the two scores separately if the two scores have a relatively low correlation. For the three carefulness tests, the correlation was approximately -0.45. The correlation between number correct and number wrong for the Mount et al. data was approximately the same: -0.46.

Davis suggested that Factor 3 reflected a carefulness attribute. Because no validity data were obtained for any of the three carefulness tests, Davis could not demonstrate any relation between scores on the tests and either training scores or operational performance. Additionally, no personality assessments were administered as part of the test battery used for the factor analysis. Such data might have allowed a more definitive interpretation of Factor 3. Without any type of supporting data, the existence of a carefulness attribute remains unresolved.

Motion Picture Testing and Research Report No. 7

Overview

The motion picture aptitude tests (Lamkin, Shafer, & Gagne, 1947) were designed primarily to assess attributes important for pilots, with a few tests developed for bombardiers. Development of most of these tests began in 1943, and, because of their long development times, many of them have no validation data. The few validation data that were collected consisted of correlations with primary pilot training completion.

Motion picture tests developed for selection can be placed into four groups: successive perception, multiple perception, judgments of motion and distance, and maintenance of orientation in space. The successive perception tests presented limited views of a moving geometrical shape, and the examinee had to guess the shape. These tests attempted to measure a visualization ability. Both the reliability and predictive validity of these tests were low and show little promise.

The multiple perception tests required monitoring five schematic instruments for instances when an indicator exceeded the tolerance value (e.g., the airspeed indicator was above the maximum allowable airspeed). These tests resemble Synwork (Elsmore, 1994), which is commercially available today. Few validity data were collected, and these show low predictive validities (r_{cbis} < 0.20).

The movement and distance estimation tests included such things as estimating the future position of an object from a limited view of its trajectory. The reliabilities of these tests were poor, and the limited data that were obtained show very low predictive validities.

The fourth group of motion picture tests assessed “the ability to maintain orientation in space” (Lamkin et al., 1947, p. 75), which appears to mean the ability to know one’s location in three-

dimensional space relative to points on the ground. The most promising of these tests was the Flying Orientation Test (CP107A). For this test, examinees were shown a movie filmed straight down from an aircraft flying over a rural area. The view of the ground appeared to be photographed through open bomb bay doors to restrict the amount of terrain visible at any given time. Periodically, the aircraft turned 90°. A "flight" consisted of four to seven legs of a constant length. At the end of the flight, the terrain faded out and was replaced by a circle with eight letters placed at equal intervals on the circle (resembling a compass rose). The examinee indicated the direction of the starting point by selecting one of eight letters. The test consisted of 25 "flights" of approximately 30 seconds each. The starting point was not visible after the first few turns, to eliminate the possibility that the examinee relied on objects or features in the terrain when making a decision. Scoring was based only on the number correct, and the test required about 30 minutes to administer. The development of this test was completed in October 1945. It may never have been administered, because no data of any type are presented.

Assessment of Motion Picture Tests

Lamkin et al. (1947) did not recommend any of the motion picture tests for continued development. Nevertheless, in the summer of 1945, several of the motion picture tests were administered with other aptitude tests in an experimental battery (Lamkin et al., 1947, p. 97). Few data from this battery were available at the time *Report No. 7* was written. However, the authors remarked that, based on the available data, the motion picture tests had low correlations with the paper-and-pencil aptitude tests. Their interpretation of this fact was that motion picture tests and paper-and-pencil tests that purportedly assessed the same attributes actually measured different ones.

The most promising motion picture test was Flying Orientation. It does, nevertheless, have one potential shortcoming: The candidate might have been able to track his movements on an imaginary two-dimensional surface instead of in three-dimensional space. This shortcoming could be addressed by including changes in altitude with changes in heading.

Psychological Research on Radar Observer Training Report No. 12

Overview

The development of radar observer selection instruments was initially performed by the National Defense Research Committee (NDRC) beginning in February 1943. In December 1944, the Army Air Forces began its own project to select and train radar observers (Hastorf, 1947, p. 11). This work was conducted by a special AAP team. The project was short-lived and was terminated after Victory in Japan Day in August 1945.

Most of the effort during this short period was devoted to conducting job analyses and developing proficiency tests for training. However, in February 1945, the AAP team fielded an experimental selection battery and conducted two validation studies. Most of the tests in both validation studies came from the November 1943 aircrew classification battery. Four of the other tests were developed by the NDRC for their research on radar operator selection (Kunsman, 1947, p. 211).[5] Another four

tests were created by the AAP team for the second validation effort. Two of these tests assessed spatial ability, and one was a job knowledge test.[6] The fourth test, Orientation to Landmarks, is the Aerial Landmarks test (CP525A) described in *Report No. 5.*[7]

The statistics reported for the first validation effort appear incomplete. The second validation study used course grades as the criterion. This second study showed low predictive validities both for the NDRC tests and the four tests created by the AAP team, with corrected biserial correlations rarely exceeding 0.10.

Assessment of Tests for Radar Observer Selection

The two new spatial tests developed by the AAP team had low predictive validities ($r_{cbis} < 0.10$). Neither of these tests is remarkable in terms of its development or content. Orientation to Landmarks (Aerial Landmarks) had a slightly higher predictive validity, $r_{cbis} = 0.13$, than the two spatial tests.[8] Nevertheless, it is still quite low. None of the tests described in *Report No. 12* are recommended for further development.

[5] Each of these three tests has an identification number identical to the type of numbers found in *Report No. 5*. However, none of them is included or even discussed in *Report No. 5*. The only description of these three tests is found in Appendix A of *Report No. 12*.

[6] The information on the two spatial tests is confusing. Pattern Identification has a different test identification number in the body of *Report No. 12* than in Appendix A. Pattern Orientation is described in the body of the report but not in Appendix A, although there is a Pattern Comprehension test in Appendix A. However, the identification numbers for Pattern Orientation and Pattern Comprehension are different. Pattern Comprehension is discussed in *Report No. 5* and has the same identification number as in Appendix A. No Pattern Orientation or Pattern Identification test is mentioned in *Report No. 5.*

[7] Why this test has two names is unclear; its development for navigators and pilots in primary training was documented in *Report No. 5* (Lacey & Niehaus, 1947, pp. 535–536), although no validation data were given. Lepley (1947) remarked that Aerial Landmarks began its development in the United States and was validated in the European theater as part of the Pathfinder Project (see *Report No. 17* below). Kunsman (1947, p. 218) implied that the process described in *Report No. 12* was a different validation effort, independent of the effort supporting the Pathfinder Project.

[8] More information is available about Aerial Landmarks under the spatial section of *Report No. 5* and in the section on *Report No. 17*.

Psychological Research in the Theaters of War Report No. 17

Overview

The selection portion of this report deals with the Pathfinder Project. This project concentrated on identifying superior pilots, navigators, and bombardiers for lead crews in bombers. Selecting good lead crew members was considered important by the operational commands because of the responsibilities of each of the lead crew members. Specifically, the aircraft commander of the lead crew commanded the bomber formation. The navigator of the lead crew had the primary responsibility for getting the formation to and from the target. The bombardier of the lead crew was responsible for identifying the target and performing the initial bombardment (Lepley, 1947, p. 17). Thus, the lead crew was very important in ensuring an effective bombing mission.

To identify superior aircrew, the AAP team adopted a two-pronged approach: They developed new selection tests, and found methods for reusing existing selection and training data. Before developing new selection tests, the team first conducted job analyses with operational commanders and flight crews in the combat theater and developed measures of combat performance. The team then developed new tests designed to be administered at the end of advanced training. One of these tests was a personality inventory assessing emotional control. Another was a "resistance to stress" test. Additionally, the team developed two new tests of spatial ability. All of these tests appear to be based on the job analyses obtained in the operational theaters, not on the job data collected from training units described in *Report No. 3* and in other reports in this series. A third spatial test, Orientation to Landmarks (Aerial Landmarks), was developed in the United States but validated for lead crew selection in the European theater.

The team also developed a selection process for lead teams that used scores from the aircrew classification battery, scores from advanced training, and instructor comments from advanced training. The validation effort primarily compared experienced lead personnel with inexperienced ones. Additionally, a limited amount of combat validation data was available from photographs taken after a strike.

Assessment of Tests Developed in Theaters of War

Few of the tests developed for lead crew identification had predictive validity data. The most promising test, Pattern Orientation, had moderate predictive validity for composite grades in lead crew training *(r* > 0.25). However, there is no indication that it was ever used operationally. As noted in the section on *Report No 12*, Pattern Orientation was originally developed for radar operators. Lead crews did not have dedicated radar operators. Consequently, Pattern Orientation might have been considered inappropriate for selecting lead crewmembers. Additionally, it never received a test designation number and is not listed in *Report No. 5.*

The most important aspect *of Report No. 17* is that it describes a secondary selection process. That is, the aircrew who experienced this selection process were already qualified for specific positions such as navigator. Four features of this process are noteworthy. The first was the need for more detailed and specific job information than that used to develop the aircrew classification battery. Collecting such data required a significant effort on the part of the team. The second was

the development of new selection tests that, logically, grew out of the new job analysis data. The third was the use of instructor comments from the advanced training courses. These comments were obtained through an unstructured collection effort but proved useful. The fourth noteworthy element of the process was the reuse of the aircrew classification battery test scores in the secondary selection process.

It is this secondary selection process that is valuable, not the individual tests that were developed for the selection process. None of the tests described in *Report No. 17* is recommended for further development.

Discussion

The 19-volume *Army Air Forces Aviation Psychology Program Research Reports* is perhaps the most comprehensive set of reports on aircrew selection ever assembled. To create the aircrew classification battery, the AAP team faced several challenges that are difficult to imagine today. One of the most serious was the lack of physical resources. The AAP team had to choose apparatus tests for the first battery based partly on the availability of parts. This problem persisted. DuBois (1947, p. 76) mentioned that some testing stations were using new apparatus tests operationally before others had even received the device. In a few cases, the shortage became so severe that no testing units were delivered on time (Melton, 1947, p. 7). Shortages also affected the development of motion picture tests. A shortage of film in 1944 and 1945 delayed the development of motion picture tests to the point that validation data were collected only for 6 of the 16 motion picture tests (Gibson, 1947, p. 2).

In addition to shortages, the AAP team constantly faced the urgent need to identify and train aircrew. The team tried using commercial tests to reduce the test development time but found that none had the necessary predictive validity. The criteria used for predictive validity estimates were always an issue, and the team spent much effort identifying better criteria than the nominal pass/fail score from primary flight training. The team also faced the constant problem of assessing performance in training and in combat operations.

The AAP research and development teams and the testing stations were structured in a rather loose manner. During the initial development of the program, a decision was made “to have a coordinated research program rather than a single, strong centralized group or a number of relatively independent units” (Flanagan, 1948, p. 12). This loose structure is reflected in some numbering and naming inconsistencies. Within a given volume, test numbering and naming is usually very consistent. However, when tests were developed by one set of AAP team members and sent to another set of team members for validation overseas, the development and validation efforts were documented in two or more volumes. In these situations, the same test might have had different names in different volumes or the same name but different identification numbers. This makes it difficult to track specific tests. In a small number of cases, one volume references another for development information or descriptive data for a specific test, but the second volume contains no information on the test in question. Generally, the cross-referencing between reports is poor.

The reports themselves seem to have been generated under considerable time pressure. All of the volumes except *Report No. 1* were published in 1947, but the prefaces often date from February 1946, approximately six months after the end of World War II. When discussing a specific test, the chapter authors frequently mentioned that the validation data were not available. In many cases, the data had been collected but the analyses were not complete at the time the chapter was written. As mentioned in a previous report (Damos, 2019), the type of paper used, the size of the fold-out pages, and the appendix formats differed widely from report to report. This again points to pressure to produce the reports quickly. Nevertheless, this series is unique in terms of both its size and the level of documentation provided.

Recommendations

The purpose of this project was to review the 19-volume *Army Air Forces Aviation Psychology Program Research Reports* to identify tests or concepts that were not pursued after World War II but have promise for today's aircrew selection system. Specifically, this project objective was to develop "systematic, prioritized recommendation of promising constructs/concepts worthy of further study for potential utility in Air Force selection and classification processes." The review generated eight recommendations, which are presented below in order of importance.

Recommendation 1: Reanalyze AFOQT Data

As discussed in the Personality and Motivation section of this report, Davis (1947b, p. 694) suggested that using only number correct or number correct with some adjustment for number wrong might result in misleading factor analyses. Although Davis had several printed tests as well as the three perceptual speed tests in his battery, he included only number correct from the printed tests in the factor analysis. For the three perceptual speed tests, Davis included both number correct and number wrong. As noted earlier, he found a factor defined only by number wrong from the perceptual speed tests. Because Davis did not include number wrong from any of the printed tests in his factor analysis, it is impossible to know if this factor is unique to perceptual speed tests or reflects a more general processing strategy or personality trait.

A reanalysis of existing AFOQT data using both number correct and number wrong might reveal the existence of a processing strategy or personality trait. However, the Air Force might choose to replicate the Davis (1947b) results by conducting a study using several perceptual speed tests in addition to the tests in the AFOQT. Mount et al. (2008) found that the number wrong on the only perceptual speed test in their battery was related to rules compliance and good citizen behavior, whereas the number correct was related to task performance. The Davis and Mount et al. results suggest that the number wrong on perceptual speed tests might reflect a distinct attribute related to cognitive style or personality.

Recommendation 2: Develop a Test Comparable to the Multiple Control Stress Test (CE210A)

This test should have at least three tasks, two of which should be tracking tasks. The test should require input from both the hands and feet, at least two displays, and cross-coupling between the tracking tasks. Auditory input may be used for a fourth test.

This test will require substantial development time and pretesting. The Air Force will need to examine strategies for task performance and learning curves. Fourier analysis probably will be necessary to determine the effects of cross-coupling on tracking performance.

Recommendation 3: Develop a New Information Integration Test

The concept underlying these tests was good, but the technology available in World War II limited the instantiation severely. An information integration test should be computerized, with multiple sources of information. The difference between this type of a test and the more common multiple-task test or divided attention test is the short-term memory load. An information integration test has a high memory load caused by extensive instructions on how to perform the task.

Recommendation 4: Develop a Memory Test Requiring the Construction of a Three-Dimensional Space with a Moving Object

The Following Oral Directions test (CP651AX) was a simple version of this type of test. A more difficult version of this test could be constructed by having the examinee change altitude each time they are attacked and decreasing the time between attacks. The difficulty of the evasive maneuvers also could be increased by requiring 45° or 135° turns rather than only 90° turns. This type of test would have high face validity and should be acceptable to the examinees.

The AFOQT has no test of either visual or auditory short-term memory. Tests assessing these features could be valuable additions to the current selection battery.

Recommendation 5: Develop a Test with Movement in Three-Dimensional Space

The AAP team attempted to determine how well an examinee could predict the future position of an aircraft from drawings of flight paths. Today, it is possible to depict movement on a computer screen in both two- and three- dimensional space. Given the slowly accumulating evidence for dynamic spatial abilities distinct from static spatial abilities (D'Oliveira, 2004), it is worthwhile to consider such a test for inclusion in a future battery.

Recommendation 6: Examine the Relation between Carefulness and Conscientiousness

The carefulness factor was defined by loadings from three perceptual speed tests. Scores on these three tests were not correlated with any personality measures. The relation of scores on these tests to any of the facets of conscientiousness is not immediately apparent and is worth exploring.

Recommendation 7: Add an Oral Memory Task with a Long-Term Memory Component

Memory for Tactical Plans (CI509BX) is an example of this kind of test. Such a test adds a long-term component that is rarely assessed in selection and classification batteries. High face validity should be achieved relatively easily. Because many memory tests present the information to be retained visually, oral presentation might assess a different facet of memory.

Recommendation 8: Increase the Number of Physics Questions in the AFOQT Physical Science Subtest

Several of the foreign air carriers who select ab initio pilots for training have added tests of mechanical principles and/or physics to their batteries. These tests add predictive validity.

References

Carretta, T. R. (1987). *Field dependence/independence and its relationship to flight training performance.* (AFHRL-TP-87-36). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Carretta, T. R. (2005). *Development and validation of the Test of Basic Aviation Skills (TBAS),* AFRL-HE-WP-TR-2005-0172. Wright-Patterson AFB, OH, Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Interface Division.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. NY: Cambridge University Press.

Cerf, A. Z. (1947). Personality inventories. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 577–621). Washington, DC: U.S. Government Printing Office.

D'Oliveira, T. C. (2004). Dynamic spatial ability: An exploratory analysis and a confirmatory study. *International Journal of Aviation Psychology, 14*, 19–38.

Damos, D. L. (2019). *Scanning the blue books*. Gurnee, IL: Damos Aviation Services, Inc.

Davis, P. C. (1947a). Mathematical tests. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 71–88). Washington, DC: U.S. Government Printing Office.

Davis, P. C. (1947b). Measures of specific traits of temperament. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 673–720). Washington, DC: U.S. Government Printing Office.

DuBois, P. H. (Ed.). (1947). *The classification program Report No. 2*. Washington, DC: U.S. Government Printing Office.

Elsmore, T. F. (1994). SYNWORK1: A PC-based tool for assessment of performance in a simulated work environment. *Behavior, Research Methods, Instruments, & Computers, 26*, 421–426.

Flanagan, J. C. (1948). *The Aviation Psychology Program in the Army Air Forces Report No. 1*. Washington, DC: U.S. Government Printing Office.

Fruchter, B. (1947a). Judgment tests. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification*

tests Report No. 5 (pp. 123–155). Washington, DC: U.S. Government Printing Office.

Fruchter, B. (1947b). Tests of set and attention. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 541–720). Washington, DC: U.S. Government Printing Office.

Gibson, J. J. (Ed.). (1947). *Motion picture testing and research Report No. 7*. Washington, DC: U.S. Government Printing Office.

Guilford, J. P. (1947). Introduction to temperament tests. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 565–575). Washington, DC: U.S. Government Printing Office.

Guilford, J. P., & Lacey, J. I. (Eds.). (1947). *Printed classification tests Report No. 5*. Washington, DC: U.S. Government Printing Office.

Hastorf, A. H. (1947). Survey of research. In S. W. Cook (Ed.), *Psychological research on radar observer training Report No. 12* (pp. 11–20). Washington, DC: U.S. Government Printing Office.

Howe, J. W., Jr., & Zimmerman, W. S. (1947). Spatial tests. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 477–510). Washington, DC: U.S. Government Printing Office.

Humphreys, L. G. (1947). Tests of intellect and information. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 45–50). Washington, DC: U.S. Government Printing Office.

Kunsman, H. F. (1947). History of radar observer selection. In S. W. Cook (Ed.), *Psychological research on radar observer training Report No. 12* (pp. 211–228). Washington, DC: U.S. Government Printing Office.

Lacey, J. I., & Niehaus, S. W. (1947). Orientation tests. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 511–539). Washington, DC: U.S. Government Printing Office.

Lacey, J. I., & Shirley, G. H. (1947). Size and distance estimation tests. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 447–476). Washington, DC: U.S. Government Printing Office.

Lamkin, H., Shafer, A. H., & Gagne, R. M. (1947). Aptitude tests. In J. J. Gibson (Ed.), *Motion picture testing and research Report No. 7* (pp. 60–98). Washington, DC: U.S. Government Printing Office.

Lepley, W. M. (Ed.). (1947). *Psychological research in the theaters of war Report No. 17*. Washington, DC: U.S. Government Printing Office.

Mashburn, N. C. (1934). The complex coordinator as a performance test in the selection of military flying personnel. *Journal of Aviation Medicine, 5*, 145–154.

Melton, A. W. (Ed.). (1947). *Apparatus tests. Report No. 4*. Washington, DC: U.S. Government Printing Office.

Mock, S. J., & Guilford, J. P. (1947). Foresight and planning tests. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 157–190). Washington, DC: U.S. Government Printing Office.

Mount, M. K., Oh, I., & Burns, M. (2008). Incremental validity of perceptual speed and accuracy over general mental ability. *Personnel Psychology, 61*, 113–139.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*.

Journal of Applied Psychology, 79, 845-851.

Raven, J. C. (1938). *Guide to using progressive matrices*. London: Lewis.

Ree, M. J. (2003). *Test of Basic Aviation Skills (TBAS) incremental validity beyond Air Force Officer Qualifying Test pilot composite for predicting pilot criteria*. San Antonio, TX: Operational Technologies Corporation.

Shirley, G. H. (1947). Memory tests. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 227–268). Washington, DC: U.S. Government Printing Office.

Thorndike, R. L. (Ed.). (1947). *Research problems and techniques Report No. 3*. Washington, DC: U.S. Government Printing Office.

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. Chicago: University of Chicago Press.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Walton, W. E. (1947). Job requirements of aircrew. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 1–11). Washington, DC: U.S. Government Printing Office.

Youtz, R. P. (1947). Part I. The fundamental elements of flying skill. In N. E. Miller (Ed.), *Psychological research on pilot training Report No. 8* (pp. 26–50). Washington, DC: U.S. Government Printing Office.

Zimmerman, W. S. (1947). Visualization tests. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests Report No. 5* (pp. 269–296). Washington, DC: U.S. Government Printing Office.