Technical Report 1183

Review of Aviator Selection

Cheryl Paullin
Personnel Decisions Research Institutes, Inc.

Lawrence Katz
U.S. Army Research Institute

Kenneth T. Bruskiewicz and Janis Houston
Personnel Decisions Research Institutes, Inc.

Diane Damos
Damos Aviation Services

July 2006

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

STEPHEN GOLDBERG
Acting Technical Director

MICHELLE SAMS
Acting Director

Research accomplished under contract
for the Department of the Army

Personnel Decisions Research Institutes, Inc.

Technical review by

David M. Johnson, U.S. Army Research Institute
Tonia Heffner, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Technical Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPC-ARI-MS, 2511 Jefferson Davis highway, Arlington, Virginia 22202-3926.

FINAL DISPOSITION: This Technical Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
July 2006	Interim	June 2004 – June 2005

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Review of Aviator Selection	DASW01-03-D-0008
	5b. PROGRAM ELEMENT NUMBER 630007
6. AUTHOR(S) Cheryl Paullin, (Personnel Decisions Research Institutes, Inc.); Lawrence C. Katz (U.S. Army Research Institute); Kenneth T. Bruskiewicz and Janis Houston (Personnel Decisions Research Institutes, Inc.); Diane Damos (Damos Aviation Services)	5c. PROJECT NUMBER A792
	5d. TASK NUMBER 308
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Personnel Decisions Research Institutes, Inc. 650 Third Ave. South Suite 1350 Minneapolis, Minnesota 55402	8. PERFORMING ORGANIZATION REPORT NUMBER Technical Report No. 493
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral & Social Sciences ATTN: DAPE-ARI-IR 2511 Jefferson Davis Highway Arlington, VA 22202-3926	10. MONITOR ACRONYM ARI-RWARU
	11. MONITOR REPORT NUMBER Technical Report 1183

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
Contracting Officer's Representative and Subject Matter POC: Lawrence Katz

14. ABSTRACT *(Maximum 200 words)*:

This report presents a review of research in the aviator selection and general personnel selection domains. That information was used to identify knowledge, skills, attributes, and other factors that should be included in a job analysis focusing on the Army aviator job. It was further used to develop a recommended strategy for an Army aviator selection battery.

15. SUBJECT TERMS
Aviator selection; selection; military; personality; psychomotor; cognitive; job analysis; review

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited		Ellen Kinzer Technical Publication Specialist (703) 602-8047

Technical Report 1183

Review of Aviator Selection

Cheryl Paullin
Personnel Decisions Research Institutes, Inc.

Lawrence Katz
U.S. Army Research Institute

Kenneth T. Bruskiewicz and Janis Houston
Personnel Decisions Research Institutes, Inc.

Diane Damos
Damos Aviation Services

Rotary-Wing Aviation Research Unit
William R. Howse, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, VA 22202-3926

July 2006

Army Project Number
633007A792

Personnel Performance and Training

REVIEW OF AVIATOR SELECTION

EXECUTIVE SUMMARY

Research Requirement:

In June 2004, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) was tasked with conducting the research and development towards a new Selection Instrument for Flight Training (SIFT). The Army's stated objectives were: 1. Develop a computer-based and web-administered selection instrument for Army flight training with emphasis upon aptitudes for Future Force aviator performance within the Future Combat Systems environment; 2. Develop an aviator selection instrument that corrects or minimizes risks associated with several deficiencies identified in the current selection instrument – the Alternate Flight Aptitude Selection Test (AFAST); 3. Develop the selection instrument so that the Army will be able to rapidly assess its current performance as a predictor, revise the instrument when necessary and adapt its application to selection for related occupational categories such as Unmanned Aerial Vehicle Operators and Special Operations Aviators; and, 4. Maximize utilization (by inclusion or adaptation) of existing tests as may be found in use or under development within the Department of Defense. The first task was to review the relevant selection literature. The overall goal of this initial task was to collect information that could be used to produce a rational recommendation for a specific selection and testing strategy for Army aviation.

Procedure:

A focused review of aviator selection research, supplemented by relevant research from the general personnel selection domain, was conducted. The review identified more than 150 potentially relevant articles. Rather than rely entirely on a narrative summary, a spreadsheet was developed to summarize information about various test batteries and to facilitate comparison of the test batteries when deriving a recommended selection strategy. From this analysis, a selection strategy for replacing the Army's current aviator selection battery was recommended. The results of this review also informed the job analysis study conducted as part of the SIFT project.

Findings:

Research clearly suggests that cognitive ability, or general intelligence (*g*), will be an important predictor of aviator performance. However, there is reason to believe that measures of the following constructs may add incremental validity beyond that achieved by a battery that reliably and accurately measures general intelligence: psychomotor skills; selective and divided attention; working memory; aviation interest/knowledge; flying experience; and, personality. The recommended selection strategy is a two-stage testing process. The first stage of testing will measure cognitive and personality/motivational traits important for the aviator job. These tests

do not require any non-standard computer peripherals and can be administered via the Internet in virtually any location with access to a desktop computer, the Internet, and a test proctor. The second stage of the test battery will include performance-based measures of psychomotor and information processing skills. These tests do require non-standard computer peripherals and may better serve the needs of Army aviation as classification instruments, for tracking selected aviators into one of the four mission platforms. Both the U.S. Navy and the U.S. Air Force currently use an aviator selection test battery that measures cognitive abilities important for U.S. Army Aviators, and one of these two batteries should be adopted for Army aviator selection. The U.S. Army also possesses two non-cognitive inventories that can be adapted for use with the Army aviator applicant population. Finally, it is recommended that a small number of new ability tests and non-cognitive scales be developed to measure abilities or traits that are not currently measured by any of the readily-accessible test batteries or non-cognitive instruments.

Utilization and Dissemination of Findings:

This product is one of many emanating from the SIFT effort. The contents of this report flow mainly into decision processes conducted internally to the project, but also documents the overall conduct of the effort. Documentation of the development of this selection instrument is necessary to provide a basis to defend the scientific and theoretical underpinnings of the test and to provide a detailed base from which revisions can be made in time. This report provides information for use in transition of the selection instrument into operation.

REVIEW OF AVIATOR SELECTION

CONTENTS

LIST OF TABLES

REVIEW OF AVIATOR SELECTION

Introduction

In June 2004, the US Army Research Institute for the Behavioral and Social Sciences (ARI) awarded the Selection Instrument for Army Flight Training (SIFT) contract to Personnel Decisions Research Institutes (PDRI). The Army's stated objectives were: 1) Develop a computer-based and web-administered selection instrument for Army flight training with emphasis upon aptitudes for Future Force aviator performance within the Future Combat Systems environment; 2) Develop an aviator selection instrument that corrects or minimizes risks associated with several deficiencies identified in the current selection instrument – the Alternate Flight Aptitude Selection Test (AFAST); 3) Develop the selection instrument so that the Army will be able to rapidly assess its current performance as a predictor, revise the instrument when necessary and adapt its application to selection for related occupational categories such as Unmanned Aerial Vehicle Operators and Special Operations Aviators; and, 4) Maximize utilization (by inclusion or adaptation) of existing tests as may be found in use or under development within the Department of Defense.

The project was divided into several tasks. This report summarizes efforts conducted in relation to Task 1: Review the existing Army aviation accession process and relevant literature. The overall goal of Task 1 was to collect information that could be used to produce a rational decision on a specific selection and testing strategy.

Overview of Existing Army Aviation Accession Procedures

A review of existing Army aviation accession procedures was conducted to provide the context for recommending a replacement for the AFAST. This included reviewing Army regulations and other documents. US Army aviators are Commissioned or Warrant Officers. Commissioned Officers primarily come from a military academy, or from a Reserve Officer Training Corps (ROTC) or Officer Candidate School (OCS) program. Civilians and enlisted personnel from any branch of the US military may apply to become an Army Aviation Warrant Officer. Prior to volunteering for aviation duty, candidates must meet standards for becoming a Commissioned Officer or a Warrant Officer in the US Army. Among other things, this includes meeting physical and medical standards, and earning a qualifying score on the relevant admission exam (Scholastic Aptitude Test or the American College Test for Commissioned Officers; Armed Services Vocational Aptitude Battery (ASVAB) General-Technical (GT) Composite for Warrant Officers).

Candidates who apply to become an Army aviator must meet additional standards beyond those described above. The selection process is rigorous and there are typically five to ten applicants for every available training seat. Selection standards are highly similar across all accession sources but the exact procedures vary to some degree, depending on whether the applicant is a Commissioned versus a Warrant Officer, the source from which he/she comes (e.g., US Army versus US Army National Guard or Reserve), and whether or not the applicant is already on active duty at the time of application. In general, all Army Aviator applicants must meet physical fitness and medical standards beyond those required to become a Commissioned or Warrant Officer, meet minimum and maximum age requirements, earn a qualifying score on

the AFAST, and be recommended by a selection board. Flight experience and post high-school coursework or degree are preferred but not required.

The following Army regulations (AR) and other documents outline selection and testing requirements:

- Selection and Training of Army Aviation Officers (AR 611-110, 14 Nov 2003)
- Aviation Warrant Officer Training (AR 611-85, 15 June 1981)
- Army Personnel Selection and Classification Testing (AR 611-5, 10 June 2002)
- Appointment of Commissioned and Warrant Officers of the Army (AR 135-100, 1 Sept 1994)
- Warrant Officer Procurement Program (Department of the Army Circular 601-99-1, 23 April 1999)
- Warrant Officer Professional Development (Department of the Army Pamphlet 600-11, 30 Dec 1996)
- Order to Active Duties as Individuals Other than a Presidential Selected Reserve Call-up, Partial or Full Mobilization (AR 135-210, 17 Sept 1999)
- Policies and Procedures for Active-Duty List Officer Selection Boards (Department of the Army Memo 600-2, 24 Sept 1999)

After candidates are selected as Army aviators, they report to Ft. Rucker, AL for training. All candidates complete an 18-week Initial Entry Rotary Wing (IERW) core training program and a two-week Basic Navigation course, followed by 12 to 20 weeks of training in a specific operational aircraft. Student aviators are assigned, or "classified" into one of four tracks for aircraft-specific training: Scout, Attack, Cargo, or Utility. Classification decisions are currently based in part on academic grades in IERW and in part on the needs of the Army. Upon completion of aircraft-specific training, aviators are assigned to a Military Occupational Specialty (MOS) that corresponds to the type of aircraft they are qualified to fly, and they begin their first operational tour as an Army Aviator.

Brief History of Aviator Selection

The prediction of aviator performance played a prominent role in the military research and development arena for most of the last century. In a review of aviator selection research, Hunter (1989) explained that this continued emphasis is a result of the expense involved in aviator training, noting that, almost without exception, aviator training is the most expensive of the training programs conducted by the military services. The US Navy estimates that the sunk costs for student aviators who fail training range from $500,000 to $1,000,000, depending on the stage at which failure occurs (Helm & Reid, 2003). According to Carretta and Ree (2000), estimates of the cost of each person who failed to complete US Air Force (USAF) undergraduate aviator training range from $50,000 (Hunter, 1989) to $80,000 (Siem, Carretta, & Mercatante, 1988). The amount approaches $500,000 per candidate by the end of flight school for US Army aviators.

Since World War I, the military services have explored the relationships between measures of a wide variety of personal characteristics and aviator performance. As early as World War I, tests of mental alertness and emotional stability were found to be predictive of aviator success (North & Griffin, 1977). Between World War I and World War II, measures of psychomotor coordination received the primary emphasis in aviator selection research. A flurry of developmental activity produced "aircraft-like controls" for use in measuring complex coordination, two-hand coordination, rudder control skills, dual-task performance, and the like. A number of these psychomotor tests, especially those of a more complex nature, were found to be valid for aviator selection. However, in the early 1950s psychomotor tests were largely abandoned as a result of persistent problems with reliability and maintainability of these electromechanical devices (Hunter, 1989).

With the advent of World War II, research on aviator selection and classification expanded to include measurement of additional abilities such as spatial orientation and the use of new testing tools (e.g., motion pictures, photographs). Much of what is known today about spatial and psychomotor abilities, as well as several other related attributes, stems from the classic Army Air Force (AAF) work (Guilford & Lacey, 1947; Melton, 1947) and the Navy's Pensacola 1000 Aviator Study (Franzen & McFarland, 1945). After the war, Fleishman and his colleagues continued psychomotor abilities research (e.g., Fleishman, 1967, 1972; Fleishman & Hempel, 1954). Researchers also investigated personality characteristics related to attrition from aviator training and/or aviator performance (Griffin & Mosko, 1977).

Within the last few decades, innovations in aviator selection and classification have centered on attributes such as multi-task performance (e.g., Griffin & McBride, 1986), division of attention (e.g., Carretta, 1987d), decision making speed (e.g., Carretta, 1988), and attitudinal and motivational traits (Foushee & Helmreich, 1986; Helmreich, Foushee, Benson & Russini, 1986). Personality also received a good deal of attention in the past two decades. Much of the early work was exploratory in nature, attempting to determine which personality traits were related to various outcomes relevant for aviators, but not necessarily guided by any particular theory of personality or aviator performance. For example, several researchers administered personality inventories that had been well established as useful for purposes other than aviator selection, including the Minnesota Multiphasic Personality Inventory (Caldwell, O'Hara, Caldwell, Stephens, & Krueger, 1993), Eysenck Personality Inventory (Bartram & Dale, 1982; Jessup & Jessup, 1971), and the Edwards Personal Preference Schedule (Fry & Reinhardt, 1969). Other researchers developed their own inventory, for example, the programmatic research conducted by the USAF that eventually led to the NEO-PI and the Self-Description Inventory (Christal, 1975; Christal, Barucky, Driskill, & Collis, 1997; Tupes & Christal, 1961).

Some of the research specifically focused on developing personality profiles for helicopter aviators (Caldwell, et al., 1993; Geist & Boyd, 1980; Harrs, Kastner, & Beerman, 1991; Howse, 1995). Another arena of increasing importance is selection of individuals to fly unmanned aerial vehicles (UAVs). For example, US Navy researchers have examined the validity of a test battery designed to measure psychomotor, multi-tasking, and visuospatial abilities in a small sample of UAV operators, with promising results (Phillips, Arnold, & Fatolitis, 2003).

This report describes the specific procedures, findings, and implications of a focused review of the aviator selection literature. As an initial step in the development of SIFT, the goal of this review was to produce a rational recommendation for a specific selection and testing strategy for Army aviation. Therefore, consideration was given to methodological limitations and obstacles in conducting selection research, as well as to the incremental validities and practical issues associated with the tests being studied.

Focused Literature Review

As noted above, aviator selection and classification research has been conducted since the 1920's and a tremendous amount has been written on this subject. This focused literature review was designed to provide a research-based foundation for a recommended selection strategy. Therefore, no attempt was made to review every aviator selection study that has ever been conducted. Rather, the focus was on key studies related to currently or recently available selection batteries, particularly those studies conducted by the US military.

The specific goals for conducting this literature review were to:

1. Review studies that delineate the knowledge, skill, ability, and other characteristics (KSAOs) important for performing the aviator job, with particular emphasis on studies that involve helicopter aviators. This information would help inform the job analysis phase of the project.

2. Review studies that focus on aviator selection batteries currently (or recently) in use by the US Air Force, US Navy, and other relevant organizations (e.g., foreign military, commercial airlines).

Literature Review Methodology

The first step in this task was to identify currently or recently available test batteries that might be viable candidates for consideration as a replacement for the AFAST and, once those were identified, to locate and summarize key research about them. This step requires consideration of a wide range of possible tests or test batteries, with the expectation that, at a later date, a number of potential candidates would be ruled out with relative ease (e.g., test batteries that cannot be computerized or ones that involve prohibitively expensive licensing fees). There was a possibility that one or more existing test batteries would be recommended as an intact entity, with minimal changes, or of recommending specific subtests from a variety of existing batteries.

Seven on-line databases were searched first, to obtain pertinent literature. These included PsychInfo, Defense Technical Information Center (DTIC), the Air Force Research Laboratory Research Archive Library, the Civil Aeromedical Institute database of technical reports, the Naval Medical Research Laboratory database of technical reports, and the archives of the Human Factors and Ergonomics Society (HFES). The HFES database covers all of the Society's publications, including the Society's bulletin and magazine. The sixth database searched was the United States Air Force Human Resources Laboratory (AFHRL, 1968-1998) Topics, hosted by the Innovation Center for Occupational Data, Applications, and Practices. All of these databases were searched using terms such as "aviator selection," "ab initio" (from the beginning), "personality," and "psychomotor." Personnel Decisions Research Institutes (PDRI) also

searched its archives for articles and technical reports related to aviator selection, based on prior work with the US Air Force, particularly in the area of Crew Resource Management (CRM).

In addition, the Damos Aviation Services (DAS) database was searched, which consists primarily of articles related to aviator selection and performance. This database currently has over 3800 entries. The earliest entry pertaining to aviator selection in the DAS library dates from 1921. It contains references to both civilian and military aviator selection, and a substantial proportion of the entries are concerned with foreign aviator selection. The DAS database covers all of the *International Journal of Aviation Psychology*, all of the *Proceedings of the International Symposium on Aviation Psychology,* and the last 19 years of *Aviation, Space, and Environmental Medicine.* Any recent materials that had not yet been entered into the database were searched by hand. Hand searches also were conducted on recently edited books that had not yet been entered into the database. Several individuals involved with aviator selection were also contacted to obtain updates on their current aviator selection research projects.

Findings from Aviator Selection Research Literature

Most of the research on aviator selection has been conducted by the military in the United States, the United Kingdom, and Norway. Some research was also published by military organizations in other countries (e.g., Israel, Turkey) and in the commercial sector. The Federal Aviation Administration (FAA) and National Aeronautics and Space Administration (NASA) have both conducted research in the arenas of cognitive and non-cognitive testing. Of most relevance for the present research is work conducted by NASA in the area of personality traits impacting aircrew performance (e.g., Helmreich, Foushee, Benson, & Russini, 1986; Musson, Sandal, & Helmreich, 2004) and work originated by the FAA's Civil Aeromedical Institute (CAMI) on a test battery called CogScreen (King & Flynn, 1995). The following sections summarize key research found in the aviator selection research literature, as well as in the general selection research literature.

General aviator selection reviews. A number of reviews of the aviator selection literature have been published (Carretta & Ree, 2000, 2003; Dolgin & Gibb, 1988; Griffin & Koonce, 1996; Hunter, 1989; North & Griffin, 1977; Ree & Carretta, 1996, 1998; Rogers, Roach, & Short, 1986; Tirre, 1997; Turnbull, 1992), including one that focuses specifically on methodological difficulties and common shortfalls associated with such research (Damos, 1996). In their review of aviator selection methods, Carretta & Ree (2000) state, "Research results point to *g* [general intelligence] as the most important underlying construct in the prediction of aviator success. Clearly, three others have been shown to be important but to a smaller degree: flying job knowledge, personality, and general psychomotor ability" (p. 31). These authors note that, "Simulation-based tests may significantly increment the validity of cognitive tests when the two approaches are used together. These results are consistent with a large-scale meta-analysis of 19 commonly used personnel selection methods across many occupations (Schmidt & Hunter, 1998)" (p. 24). Regarding personality measures, Carretta and Ree comment that a great deal of research has been conducted in this area, with contradictory results. They go on to say that organizing the results according to the Big Five personality variables of Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness (Norman, 1963; Tupes &

Christal, 1961) would likely be enlightening, but has not (yet) been done in the aviator selection arena.

Griffin and Koonce (1996) also wrote a comprehensive review of aviator selection, with particular emphasis on measures of psychomotor skills. They review numerous research studies showing that several types of predictor measures are valid for predicting aviator performance, including:

- aptitude (cognitive ability);
- psychomotor skills;
- work simulation;
- divided attention (or multiple-task performance);
- flying experience; and,
- biographical information.

According to these authors, uncorrected, zero-order correlations for psychomotor skills are in the .30 to .40 range and multiple regression correlations are in the .50 range in research studies involving continuous criterion measures such as instructor check/flight ride ratings. With regard to measures of psychomotor skills, the authors concluded,

> Automated versions of vintage psychomotor tests (developed in the 1930s and 1940s) seem to be as predictive of military aviator/aviator performance today as in the past. The use of computers may have enhanced the predictive power of the psychomotor tests by making their functioning dependent on digital electronic circuitry, rather than analog electromechanical devices, resulting in more reliable performance measurement. The psychomotor tests receiving the most attention today are the CCT [complex coordination test] and the THCT [two-hand coordination test], originally developed by Mashburn and colleagues before World War II (Mashburn, 1934). These tests were significant predictors of USAF and Navy pass-fail criteria in the past, and automated versions are predictive today. However, the tests are better predictors of normally distributed, continuous criteria such as flight grades and number of flight hours for the Navy and check rides and advanced training ratings for the USAF [than of traditional pass-fail] (p. 143).

Tirre (1997) made a useful distinction between two different approaches to aviator selection:

- Basic attributes – In this approach, the test battery measures specific attributes that are assumed to underlie aviator performance. Examples of this approach include the USAF's Air Force Officer Qualifications Test (AFOQT) and Basic Aviator Test (BAT; Carretta, 1987a).
- Learning sample (simulation) – In this approach, the test battery simulates tasks performed in flight, with varying degrees of realism. An example is the Canadian Automated Pilot Selection System (CAPSS).

Each approach has advantages and disadvantages, many of which are outlined by Tirre. The basic attributes approach has a long history outside aviator selection and is generally less costly and time-consuming to develop and administer than the learning sample approach. In fact, it has only been possible to use the learning sample approach widely and effectively with the advent of powerful desktop computers. The learning sample approach offers the advantage of dynamic (as opposed to static) measurement of cognitive processing skills and often involves measures that appear very realistic to test-takers. With either approach, the reliability and validity of the measurement tool depends critically on how carefully it was developed.

Obstacles and Issues in Conducting Aviator Selection Research

There are several obstacles to conducting research studies in the aviator selection domain, many of which have been recognized for a long time and many of which are exceedingly difficult to overcome. These issues have been described in several of the preceding reviews (e.g., Carretta & Ree, 2000; Damos, 1996), and the most important ones are summarized below. When reviewing the literature, it became clear that some researchers recognized these obstacles and acknowledged how their study results and conclusions were likely impacted; many others did not.

Training Performance as a Criterion Measure

The criterion measure in aviator selection research studies is almost always a measure of training performance. While training performance is clearly an important outcome measure, it certainly is not the only outcome variable of interest. Unfortunately, it is exceedingly difficult to obtain reliable and accurate measures of aviator performance after training. The reliance on training performance as a criterion measure is particularly problematic because researchers are typically unable to differentiate various types of "failure." Different abilities or traits may underlie different types of failure, but the pattern of relationships will be difficult or impossible to detect if there is no way to identify and code the reason(s) for failure.

The reliance on training outcome measures is also problematic when attempting to evaluate the validity of predictor measures that would not necessarily be expected to predict training performance (e.g., personality measures). Research conducted as part of the US Army's Project A shows that measures of cognitive ability predict declarative knowledge and technical components of performance (McCloy, Campbell, & Cudeck, 1994) while measures of non-cognitive characteristics predict motivational aspects of job performance (Campbell, Hanson, & Oppler, 2001; McCloy, Campbell, & Cudeck, 1994) and contextual performance (Borman, Penner, Allen, & Motowidlo, 2001; Campbell, Harris, & Knapp, 2001; Campbell & Knapp, 2001). While motivational factors certainly play a role in training performance and most students are highly motivated to succeed, the type of training criterion *measures* typically used in aviator selection research do not separate technical performance and motivational aspects of performance. Thus, training criterion measures are likely more heavily weighted toward academic and technical aspects of performance (e.g., flight instructor ratings, grades, pass-fail status) and less heavily on motivational aspects of performance.

Statistical/Methodological Issues

Aviator selection research is plagued by a number of statistical and methodological issues. Some of them are extremely difficult, if not impossible, to overcome.

1. *Using predictor or criterion measures of low or unknown reliability.* In many cases, the reliability of predictor and criterion measures is not reported and may be quite low, particularly in the case of criterion measures. Thus, the impact of unreliable measurement on the outcomes of the study cannot be evaluated.
2. *The most common criterion measure is a dichotomous variable – pass-fail status at the end of training.* When working with dichotomous criterion variables, the highest possible value of a correlation between any predictor measure and that criterion variable depends on the distribution of the dichotomous variable. The maximum possible value of the correlation is lower the more the distribution varies from a 50-50 split. For aviator training pass-fail status, the pass-fail distribution is usually much more extreme than 50-50. It is possible to correct the correlation coefficient for dichotomization, and some researchers did this. It is important to note that, while pass-fail performance in training is impacted by the attitudes and skills of student aviators, it is also impacted by the policies of aviator accession and training organizations. When there is a strong need for aviators, for example during war, there is strong pressure to ensure that virtually all students will pass training. In addition, most aviator training programs make every effort to ensure that most students pass training because it is very costly to fail a candidate after several weeks of expensive training.
3. *Aviator selection research is based on a highly selected and homogeneous population.* Before they begin an aviator training program, all applicants have been extensively screened, including meeting a required minimum score to enter the military, meeting a required minimum score on an aviator aptitude battery, meeting education requirements, and/or earning strong, positive evaluations and recommendations from a superior officer or a selection board. The samples used in most aviator selection research are also typically highly homogeneous in terms of race and gender. Screening occurs in multiple stages, with each stage serving to further restrict the sample relative to the general population. Correlations can be corrected for some types of range restriction, but there is disagreement about the extent to which such corrections should be made. Damos (1996) argues that it is not appropriate to make such corrections because aviators will never be selected from an unrestricted sample. In addition, some types of restriction cannot be corrected for, including the demographic composition of the sample.
4. *Failure to correct for capitalization on chance.* A number of researchers in the aviator selection domain have used regression techniques to evaluate the validity of a test battery, without recognizing or correcting for the fact that such techniques capitalize on chance variations present in their sample. The reported multiple correlation may not generalize to a new sample, especially if the original sample was not very large.

5. *Small sample sizes.* In some aviator selection research, the sample is very small. This means there may have been very little power to detect significant relationships even if they did exist.

6. *Measurement method is confounded with measurement target.* As Carretta and Ree (2000), Hough (2001), and others have noted, in some research studies, measurement *method* (e.g., biodata or personality inventory) is confounded with the measurement *target* (i.e., KSAOs). In some cases, there is a close correspondence between measurement method and measurement target. For example, "psychomotor tests" virtually always measure one or more psychomotor abilities, and typically very little else. In contrast, the "biodata" or "personality" measurement method can be used to target leadership tendencies, conscientiousness, stress tolerance, psychopathology, motivation, or other KSAOs. Summarizing findings across all biodata inventories or all personality inventories tells us little about which underlying traits are more and less predictive of aviator performance. The situation is worsened by the fact that not all biodata and personality inventories measure the same targets. Thus, across studies, there may be a great deal of variation in the extent to which relevant and irrelevant KSAOs are measured.

Low Base Rate

Predictor variables. Some tests are designed to identify applicants, aviator trainees, or experienced aviators who have a severe psychopathological problem or a neurological deficit. Tools such as CogScreen, dichotic listening tests, and the MMPI have been used for this purpose. King and Flynn (1995) describe CogScreen as "a self-administered screening tool, in which the subject uses a light pen on a cathode ray tube monitor. CogScreen may be superior to traditional neuropsychological testing in determining cognitive deficits after a central nervous system injury or dementing disease . . . . The CogScreen is very sensitive to the nuances of neuropsychological functioning and can be administered in a group setting" (p. 954). No validity studies for CogScreen could be located. The USAF explored the possibility of using CogScreen for aviator medical screening, but it was never used operationally for aviator selection. The US Navy is currently including CogScreen, or a variation of it, in their ongoing studies to enhance aviator selection.

Severe psychopathology and neurological deficits are rare in the general population, and are even rarer in the highly-selected population of aviators (including applicants and trainees). While it may be exceedingly important to identify individuals in the aviator population who might or will experience these problems, doing so is literally like "looking for a needle in a hay stack." The low base rate for these problems makes it extremely difficult to show a statistically significant relationship between test scores and outcome measures, even if the test is valid. In past research, the failure to find significant correlations for these types of tests was sometimes inappropriately generalized to *all* tools of a particular type, for example, all personality inventories. Callister, King, Retzlaff, and Marsh (1999) point out, "Testing for psychopathology has been shown to be of limited value in the assessment of the highly-functioning aviator population. On the other hand, measures of normal personality characteristics have been shown to be useful in a variety of settings and populations" (p. 885).

Criterion variables. As noted above, the most common criterion measure is pass-fail status at the end of aviator training, and the base rate for failure is typically low. The low base rate issue becomes even more extreme when researchers attempt to categorize failure according to type or reason, for example, failure due to lack of technical competence versus failure due to attitudinal problems, or when the training failure rate is mandated by policy to be extremely low.

Factor-Analytic Work in the Aviator Selection Research Literature

In spite of the aforementioned limitations, selection test developers have continued to search for measures that might predict aviation performance. Accordingly, researchers have factor analyzed scores on several aviator selection batteries to uncover which constructs yield incremental predictive validity. Most of the work was conducted by USAF researchers. Several of these studies are summarized below.

Carretta and Ree (1997a) administered the Armed Services Vocational Aptitude Battery (ASVAB) and 17 psychomotor tests to enlisted USAF personnel (n = 429). They summarized their findings as follows:

> Confirmatory factor analysis yielded higher-order factors of general cognitive ability (g) and psychomotor/technical knowledge (PM/TK). PM/TK was interpreted as Vernon's (1969) practical factor (k:m). In the joint analysis of these batteries, g and PM/TK each accounted for about 31% of the common variance. No residualized lower-order factor accounted for more than 7%. PM/TK influenced a broad range of lower-order psychomotor factors. The first practical implication of these findings is that psychomotor tests are expected to be at least generally interchangeable. A second implication is that the incremental validity of psychomotor tests beyond cognitive tests is expected to be small (p. 165).

Ree & Carretta (1992) conducted a similar study, using the ASVAB and three psychomotor tests from the Basic Aviator Test (Carretta, 1987a). The sample was 354 USAF enlisted recruits. They found that the two types of tests correlated with each other, with average correlations in the .30's (corrected for range restriction, but not for test unreliability). They also found that, as expected, there was a large first factor, which they labeled "psychometric g," and that both the ASVAB and the psychomotor tests loaded on it. Confirmatory factor analyses revealed that both a seven-factor and a nine-factor model fit the data equally well. The more parsimonious seven-factor model includes psychometric g and a higher-order general psychomotor factor which accounted for 57% and 9% of the total variance respectively. Other factors included 1) Verbal-Technical (accounted for an additional 8% of the variance), 2) Non-technical General Knowledge (10%), 3) Time-Sharing (4%), 4) Two Hand Coordination (7%), and 5) Complex Coordination (5%).

Carretta and Ree (1998) compared the factor structure of the ASVAB with the factor structure of the AFOQT. The factor structure for each test battery was derived in a different sample of USAF personnel, because the two batteries differ in difficulty level and intended audience (with the ASVAB being taken by all Air Force applicants and the AFOQT being taken by Flight Officer applicants). The authors conclude "The AFOQT is comprised of five lower-order factors: verbal, math, spatial, aircrew, and perceptual speed which accounted for 20% of the total variance, and g in hierarchical position accounted for 41% of the total variance.

Compared with the ASVAB, the AFOQT was less saturated [with *g*] but had more common factors and had a greater proportion of its variance associated with common factors" (p. 12).

Carretta, Retzlaff, and King (1997) compared the AFOQT and the Multidimensional Aptitude Battery (MAB). The MAB is a broad-based test of intellectual ability patterned after the Wechsler Adult Intelligence Scale but designed for group administration. The sample in this study was approximately 2,200 USAF aviator candidates. A joint factor analysis of the AFOQT and the MAB revealed that each battery had a hierarchical structure. The correlation between the higher-order factors from the two batteries was .981, indicating that both measured the same thing, which these authors conclude is general intelligence (*g*).

Ambler and Smith (1974) analyzed data for the seven tests of the Guilford-Zimmerman Aptitude Survey, the Hidden Figures Test, and four subtests from the US Navy-Marine Corps aviation selection battery [1) Aviation Qualification, which includes reading, math, and science questions related to a typical college experience; 2) Mechanical Comprehension; 3) Spatial Apperception; and 4) Biographical Inventory]. Scores were available for approximately 1,700 aviation trainees (presumably all male, given that the study was published in 1974). The researchers factor analyzed the subtest scores in the total sample and in various subsamples and found that six factors appeared consistently across samples, which they labeled Mechanical, Spatial Manipulation, Perceptual Flexibility, Verbal Intelligence, Numerical Intelligence, and Flight Motivation.

Martinussen and Torjussen (1998) factor analyzed scores on a multi-aptitude test battery used for aviator selection into the Norwegian Air Force. The battery is administered in a multi-stage process. Stage 1 includes 12 subtests intended to measure General Intelligence, Technical Comprehension, and Spatial Ability. Stage 2 includes seven subtests intended to measure Simultaneous Capacity and Orientation Ability. Finally, Stage 3 includes a personality inventory called the Defense Mechanism Test (DMT) which is described as a measure of psychodynamic defense mechanisms and was developed for use in selecting persons into high-risk professions. Very little information is provided about any of the subtests.

The authors randomly selected 450 applicants from the applicant pool who had Stage 1 and Stage 2 scores, and factor-analyzed the scores using Principal Component Analysis with Varimax rotation. The tests included in each stage were factor-analyzed separately. Three factors, labeled Mechanical Comprehension and Spatial Ability, Verbal Ability, and Numerical Reasoning accounted for 61% of the variance in the Stage 1 tests and three factors, labeled Spatial Ability, Time Estimation, and Perceptual Speed and Coordination, accounted for 62% of the variance in the Stage 2 tests.

In summary, factor analyses of several aviator selection batteries suggest that it is possible to derive a hierarchical general intelligence factor, with sub-factors related to verbal ability, numerical ability, mechanical ability, spatial ability, and perceptual speed/flexibility. A general psychomotor factor, with some specific sub-factors also appears when the test battery explicitly contains psychomotor tests.

The factor-analytic work in the aviator selection domain is consistent with research conducted on the structure of human abilities that is not entirely based on military or aviator test

data (e.g., see Fleishman & Mumford, 1988; Lubinski & Dawis, 1992; McHenry & Rose, 1988; Russell, Reynolds, & Campbell, 1994). It is worth noting that, with the exception of one subtest in the US Navy-Marine Corps selection battery (Ambler & Smith, 1974), the aviator selection batteries included in the factor-analytic studies described above did not include measures of non-cognitive traits. It is not particularly surprising, then, that no underlying non-cognitive factors were found.

Models of Skill Acquisition

This section examines more closely the hierarchical general intelligence factor derived by the factor analyses described above. Specifically, the question, "How does intelligence relate to skill acquisition during flight training?" is addressed.

Ackerman (1987; 1988; 1990) developed a model of skill acquisition that is applicable to the development of piloting skills. The theory is founded on the concept of attentional resource allocation, that is, the amount of attentional resources required by various tasks at various points in time, and the amount of attentional resources that individuals can bring to bear in any given situation. Ackerman's model divides skill acquisition into three broad phases, with a corresponding type of ability that is the primary predictor of performance within each phase. In Phase I, the primary learning task is to comprehend the new task. Declarative knowledge and general intelligence are the primary predictors of performance in this phase. In Phase II, the primary learning task involves integrating the cognitive and motor processes required to perform the task. In this phase, knowledge compilation and perceptual speed are the primary predictors of performance. In Phase III, task performance becomes proceduralized (or automatic), and thus requires fewer attentional resources. Procedural knowledge and psychomotor abilities are the most important predictors in this phase. Tasks vary in the extent to which they can be proceduralized. In Ackerman's terminology, tasks that can become proceduralized are called consistent tasks; those that cannot become proceduralized are called inconsistent tasks.

According to Ackerman's theory, general cognitive ability is expected to be most important during the early stages of skill acquisition for all tasks *and* to remain important for inconsistent tasks. Processing speed is expected to be most important during intermediate stages of learning for any task. Psychomotor skills will become increasingly important as a task becomes better-learned, but may only outstrip cognitive ability in importance for inconsistent tasks. Keil and Cortina (2001) found confirmatory evidence for the relationship between cognitive ability and performance on consistent and inconsistent tasks but did not find support for the relationship between perceptual speed and psychomotor skills and consistent and inconsistent tasks. Additional research by Ackerman and colleagues shows that both ability and non-ability factors (e.g., personality, vocational interests, motivation, and self-concept) play a role in determining performance on complex (inconsistent) tasks (Ackerman, Kanfer, & Goff, 1995; Ackerman & Woltz, 1994).

Ree, Carretta, and Teachout (1995) developed a causal model to explore the role played by general intelligence (*g*) and prior knowledge of flying on performance during aviator training. The measures of *g* and prior flying knowledge were based on AFOQT composite and subtest scores collected at the time of application to flight training. Criterion measures included measures of job knowledge (academic classroom performance) and work samples (check ride performance) collected at various points during a 53-week training program. When the model

was tested in a large sample of USAF aviator trainees (n = 3,428 males), the authors found that g directly influenced the acquisition of flight knowledge both prior to and during training and indirectly influenced work sample performance through the acquisition of job knowledge. Prior knowledge of flying had almost no influence on acquisition of job knowledge during the academic portions of aviator training, but directly influenced performance on early work sample measures. Early work sample performance was very strongly related to later work sample performance. Carretta and Ree (1997b) tested the same model in a sample of male USAF aviators (n = 3,369) and in a small sample of female USAF aviators (n = 59). The basic model was supported and appeared to work similarly for males and females, although the female sample was too small to draw any strong conclusions.

Evidence of Predictive Validity for Flight Training Performance

An enormous number of validation studies have been conducted in the aviator selection domain – too many to cover in this report. Fortunately, several meta-analyses focusing on the validity of selection tests have been published. In all the studies, measurement *method* is confounded with measurement *target (*KSAOs) to at least some degree. This section describes meta-analyses addressing validity evidence.

Damos (1993) Meta-Analysis

The first meta-analysis of aviation performance predictors was published by Damos in 1993. She meta-analyzed 12 studies that involved a single-task performance-based measure, for example, tracking or dichotic listening, and 14 studies that involved multiple-task performance-based measures, that is, two or more single-task measures administered simultaneously, such as tracking plus dichotic listening. The mean correlation (uncorrected) between single-task performance and flight grades was .18 (n = 5,378); the correlation between multiple-task performance and the same criterion was .23 (n = 6,920). Moderator analyses suggested that the level of validity for multiple-task performance-based measures depended on the type of sample (military versus civilian) and level of flight experience (students versus fully-trained aviators), with higher validity in studies with a civilian sample or with a fully-trained aviator sample.

Hunter and Burke (1994) Meta-Analysis

The second meta-analysis was published by Hunter and Burke in 1994.[1] They reviewed 200 studies published between 1940 and 1990 that involved aircrew selection. Sixty-nine studies contained one or more usable validity coefficients, and the authors located or derived 468 validity coefficients from these studies. It is worth noting that studies reporting only a composite score based on a multi-aptitude test battery were excluded from the meta-analysis. The majority of the validity coefficients were based on studies conducted in the US (77%), involving a military sample (94%), and/or a sample that was training to fly fixed-wing aircraft (86%). Most of the studies used dichotomous pass-fail criterion measures (84%) which, as noted above, places

[1] An earlier version of this meta-analysis was also published in Hunter and Burke (1992). The general findings are the same in the two versions, but the specific values cited for various predictor types is not exactly the same.

a ceiling on the maximum possible correlation, with a lower ceiling to the extent the criterion distribution departs from a 50-50 split (as is likely the case in virtually all of the studies).

Hunter and Burke (1994) categorized each validity coefficient according to one of 16 predictor types. They then applied bare-bones meta-analytic procedures (Hunter & Schmidt, 1990). Table 1 is adapted from a table of results published in Hunter and Burke (1994). It shows, for each predictor type, the mean sample-weighted validity (uncorrected), the percentage of variance explained by sampling error, and the lower bound for the 95% confidence interval.[2] The predictor type with the highest mean sample-weighted validity is "Job Sample." The authors do not describe this predictor type, but one might speculate that it includes flight simulation tests. Mechanical ability, gross dexterity, reaction time, biodata inventory, and information (General or Aviation) predictors also showed relatively high validity and a confidence interval that did not include zero. Recall that "biodata" is a measurement method. There is no way of determining, from this meta-analytic review, what KSAOs were measured.

After conducting the bare-bones meta-analysis, Hunter and Burke applied two validity generalization decision rules: 1) Does sampling error account for more than 75% of the variance in observed validities? and 2) Does the 90% credibility limit include zero? Answering "no" to the first decision rule allows one to conclude that validity is generalizable across samples and settings. Answering "no" to the second decision rule allows one to conclude that the true validity in the population is greater than zero. None of the predictor types included in this meta-analysis met the first decision rule, but several met the second. For these predictor types, it is reasonable to believe that the true validity is greater than zero in any setting or sample, but the level of validity may vary from one setting or sample to another: Quantitative Ability; Spatial Ability; Mechanical; Aviation Information; General Information; Gross Dexterity; Perceptual Speed; Reaction Time; Biodata Inventory; and, Job Sample.

[2] Hunter and Burke (1994) claim that, in keeping with decision rules established by Hunter & Schmidt (1990), they calculated and used the 90% credibility limit, rather than the 95% confidence interval. In their table of results, however, they report the 95% confidence interval.

Table 1

Hunter and Burke (1994) Meta-analytic Results for Various Predictor Types

Predictor Type	# of Correlations	Total Sample Size	Mean *r*	% Variance Explained by Sampling Error	95% CI Lower Bound
General Ability	14	8,071	.13	21%	-.05
Verbal Ability	17	22,841	.12	6%	-.09
Quantitative Ability	34	46,884	.11	28%	.01
Spatial Ability	37	52,153	.19	14%	.05
Mechanical	36	42,418	.29	8%	.11
General Information	13	29,951	.25	4%	.06
Aviation Information	23	25,295	.22	12%	.06
Gross Dexterity	60	48,988	.32	13%	.15
Fine Dexterity	12	2,792	.10	45%	-.09
Perceptual Speed	41	33,511	.20	19%	.05
Reaction Time	7	10,633	.28	16%	.16
Biodata Inventory	21	27,004	.27	6%	.07
Age	9	13,810	-.10	11%	-.25
Education	9	6,163	.06	12%	-.16
Job Sample	16	2,814	.34	37%	.19
Personality	46	22,486	.10	11%	-.16

Notes.

1. Mean *r* is weighted by sample size, but has not been corrected for any other artifacts.
2. When analyzing the data, validity coefficients were reflected for predictor types that would be expected to show a negative correlation with the criterion variable, that is, those involving measures of speed. Thus, in the table above, positive correlations indicate that better performance on the predictor is associated with better performance on the criterion measures.

According to Hunter and Burke (1994), the following predictor types may show non-zero validity in some settings or samples:

- General Ability
- Verbal Ability
- Fine Dexterity
- Age
- Education
- Personality – Recall that "personality" is a measurement method. Across studies, some of the personality scales likely were expected to show a negative correlation with criterion performance (e.g., Anxiety), while other scales were likely expected to show a positive correlation (e.g., Self-Confidence). For still other personality scales, there likely was no clear *a priori* expectation about the direction of the correlation (e.g., Risk-Taking). One might argue that averaging across all the different types of scales does not provide an accurate representation of the true level of validity that might be achieved by measures of specific personality traits.

Hunter and Burke conducted moderator analyses for a subset of the predictor types for which there were sufficient data. They examined four possible moderators: 1) time period in which the study was conducted (1940-1960 versus 1961-1990), 2) nationality of the study sample (US versus other), 3) service branch (Air Force versus other), and aircraft type (fixed-wing versus rotary-wing). The most consistent finding was that the time period in which the study was conducted moderated the validity of several predictor types, with lower mean validity in more recent studies. The authors speculate that the decline in validity over time could be due to reduced variability in the applicant pool, more extreme splits on dichotomous criterion measures (e.g., farther away from a 50-50 split in the proportion of trainees who pass versus fail UPT), or changes in the nature of aviator training. The other moderator variables, at least as coded in this meta-analysis, provided very little explanatory power.

Martinussen (1996) Meta-Analysis

The third meta-analysis was published by Martinussen in 1996. She conducted a standard computerized literature search and also made a special effort to collect unpublished validation studies focusing on military aircrew selection from researchers in NATO countries. Studies that did not report the magnitude of nonsignificant correlations or only reported corrected correlations were excluded. (Hunter and Burke do not say how they handled such studies). Martinussen reports that she reviewed 134 studies, and located 66 independent samples in 50 studies that met her criteria for inclusion. Fifty percent of the studies were conducted in the United States. Most samples involved military aviators, with the bulk of those belonging to the Air Force. Two-thirds of the studies involved fixed-wing aviators and 21% involved rotary-wing aviators (12% did not specify the type of aircraft). Twenty (40%) of the studies were unpublished material. All of the studies used performance during aviator training as the criterion variable – dichotomous pass/fail status, instructor ratings, or course grades. While the distribution of study types is similar to that described by Hunter and Burke (1994), comparison

of the reference lists reveals very little overlap in the studies included in each review. In fact, fewer than 20 studies appeared in both meta-analyses.

Martinussen categorized each predictor measure into one of nine measurement *methods* (predictor types). Each is described below:

1. *Cognitive* includes all tests designed to measure a specific type of cognitive ability (e.g., mechanical, spatial, verbal, quantitative).
2. *Intelligence* includes tests specifically designed to measure global intelligence.
3. *Psychomotor/Information Processing* includes all tests involving apparatus or a computer. Obviously, this could encompass several different types of ability measures (e.g., psychomotor skills, reaction time, etc.)
4. *Aviation information* includes tests with questions about aviation. Martinussen points out that most psychologists interpret such tests as measures of motivation to become an aviator.
5. *Biographical inventories* collect background information about applicants, and then summarize the information according to a total score. Although Martinussen does not comment on the nature of the inventories, it is likely that many of them were empirically-scored.
6. *Personality* tests include a variety of personality inventories. The data are not organized according to personality trait but, unlike Hunter and Burke (1994), Martinussen did attempt to take the expected relationship between the underlying scale and the criterion variable into account by reflecting the sign of the correlation, if needed, based on information in the original study. In cases where no expectation about the direction of the relationship could be derived from the original study, Martinussen coded the absolute value of the correlation, in effect making it positive. This has the overall effect of inflating the mean validity coefficient.
7. *Combined index* was used when a validity coefficient was reported only for a combination of predictor measures. Martinussen does not report how, or if, she took account of the fact that such measures may capitalize on chance, for example, if they were created using a regression procedure. (Hunter and Burke excluded these studies.)
8. *Academics* includes school grades or tests that measured mathematical or language proficiency.
9. *Training experience* includes measures of flying performance prior to selection into the training program that was the focus of the study. It is not clear if these included self-reported or verified, objective measures of prior flight hours/performance, or both.

Table 2 shows the number of correlations, total sample size, mean sample-weighted correlation (observed and corrected for dichotomization), percent variance explained by sampling error, and 90% credibility limit for each measurement method. Using decision rules similar to those applied by Hunter and Burke (1994), Martinussen suggests that the mean validity of the Academics measurement method ($r = .15$) is likely to generalize across samples and

settings, given that 70% of the variance in observed validity is explained by sampling error. For the remaining measurement methods, it appears that there may be moderator variables that impact the level of validity across settings and samples, but the 90% credibility limit is greater than zero for all but two of them (biographical inventory and personality).

Martinussen (1996) also found a negative correlation between year of study publication and validity coefficients for each type of predictor measure except Training Performance. This is consistent with the finding of a decline in validity across time reported by Hunter and Burke (1994). She also conducted several moderator analyses. Of most interest for the present effort was her finding of a significant difference in the mean validity of two measurement methods – (general) intelligence and training experience — depending on type of aircraft. General intelligence tests showed higher validity in samples of rotary-wing aviators (mean uncorrected $r = .27$) than in samples of fixed-wing aviators (mean uncorrected $r = .11$).

Table 2

Martinussen (1996) Meta-analytic Results for Various Measurement Methods

Measurement Method	# of Correlations	Total Sample Size	Mean *r*	% Variance Explained by Sampling Error	90% Credibility Limit
Cognitive	35	17,900	.22 (.24)	12%	.07
Intelligence	26	15,403	.13 (.16)	18%	.03
Psychomotor/Info Processing	29	8,522	.20 (.24)	28%	.10
Aviation Information	16	3,736	.22 (.24)	46%	.14
Personality	21	6,304	.13 (.14)	24%	.00
Biographical Inventory	13	11,347	.21 (.23)	4%	.00
Combined Index	14	5,362	.31 (.37)	13%	.19
Academics	9	4,267	.15	70%	.11
Training Experience	10	5,806	.25	7%	.07

Note. Mean *r* is weighted by sample size. The value enclosed in parentheses is the sample-weighted mean *r* corrected for criterion dichotomization.

In contrast, training experience showed higher validity in samples of fixed-wing aviators (mean uncorrected r = .35) than in samples of rotary-wing aviators (mean uncorrected r = .12). The latter finding may be due to the fact that individuals who pursue a private pilot's license prior to entering a formal aviator training program are more likely to do so in a fixed-wing aircraft. As a consequence, the training experience may more directly transfer to, and thus positively affect, performance in fixed-wing aviator training than in rotary-wing aviator training. This finding is also consistent with anecdotal evidence that "too much" prior experience or training in fixed-wing aircraft can be detrimental when learning to fly a rotary-wing aircraft.

Martinussen and Torjussen (1998) Meta-Analysis

The fourth meta-analysis, conducted by Martinussen and Torjussen (1998), focused exclusively on a test battery used for aviator selection into the Norwegian Air Force (NAF). Four studies were included, with two to five independent samples for each of 19 subtests included in the test battery. Sample sizes ranged from 244 to 977 per subtest. In all four studies, the test battery was used in the aviator selection process so there was direct restriction of range on the subtest scores. Furthermore, spatial abilities were measured in each of two successive stages of the battery, albeit with different tests. As a consequence, the final sample was highly restricted in terms of spatial ability. Criterion measures were based on training performance, primarily pass/fail status, but also instructor ratings and course grades.

Out of 19 subtests, 10 showed a 90% credibility limit greater than zero. The mean uncorrected validity was lower than .20 for all but two of them – Aviation Information (mean uncorrected r = .21) and Instrument Comprehension (mean uncorrected r = .26), both of which were administered in the first stage of testing. Martinussen corrected the validities for dichotomization of the criterion measure (when appropriate), but did not correct them for range restriction. The corrected validities are consistently somewhat higher. One can only speculate how high they might be if corrected for range restriction as well.

Interestingly, this is the only study in which the 90% credibility limit was greater than zero for a personality measure, although the mean validity was still low and consistent with the level reported in other meta-analytic reviews (mean r = .06 and .12 for two non-independent scoring methods used within the same inventory). According to Martinussen and Torjussen, the personality inventory – the DMT – measures psychodynamic defense mechanisms, and was specifically developed to select personnel for high-risk professions. The Norwegian Air Force used it as a post-selection screening device for individuals who had already been selected into aviator training.

Summary of Meta-Analytic Validation Studies

As noted above, there was very little overlap in the studies included in three of the meta-analytic reviews. (The obvious exception is that all four of the studies included in the Martinussen and Torjussen (1998) review also appeared in Martinussen's (1996) broader review.) Only four of the studies reviewed by Damos (1996) appear in the Hunter and Burke (1994) citation list, and only one appears in the Martinussen (1996) citation list. Fewer than 20 references appear in both the Hunter and Burke (1994) and Martinussen (1996) reviews. Different authors also categorized the predictor measures differently, making it difficult to

compare the results from different reviews. Nevertheless, the following summary statements can be made:

- Global intelligence tests showed about the same, relatively low level of validity in the two meta-analyses in which they were included (mean uncorrected $r = .13$), with support for validity generalizability in one study but not in the other.
- The validity of specific cognitive ability tests seems to vary depending on the type of ability being measured but tends to be higher than that of more global measures of intelligence. This statement is supported by the mean uncorrected validity of .22 for the cognitive measurement method, as opposed to the mean uncorrected validity of .13 for the global intelligence measurement method, as reported in Martinussen (1996). It is also supported, to some degree, by Hunter and Burke's finding that the mean uncorrected validities for two specific cognitive ability predictors types, Spatial ($r = .19$) and Mechanical ($r = .29$), are higher than that for the global intelligence predictor type ($r = .13$). However, the mean uncorrected validity of verbal and quantitative ability predictor types reported by Hunter and Burke ($r = .12$ and .11, respectively) is about the same as that reported for the global intelligence predictor type. This finding may be at least partially due to higher content overlap between global intelligence and verbal and quantitative ability tests than between global intelligence and spatial or mechanical ability tests.
- There is some evidence that Mechanical ability tests are among the more valid measures of performance during aviator training, as evidenced by the mean uncorrected validity of .29 in the Hunter and Burke (1994) meta-analysis and the mean uncorrected validity of .26 for the Instrument Comprehension subtest in the Martinussen and Torjussen (1998) meta-analysis. (Factor analyses of the Norwegian Air Force test battery suggested that the Instrument Comprehension measures both mechanical and spatial abilities.)
- Aviation Information tests showed about the same level of validity in the three meta-analyses in which they were included – about .22 (uncorrected).
- The biographical inventory measurement method showed a relatively high mean validity in the two meta-analyses in which it was included (mean uncorrected $r = .27$ and .21, respectively). However, sampling error explained very little of the variability in validity estimates across studies, suggesting that there are other factors that impact the validity of such inventories. One of the most important factors may be the extent to which the inventory was designed to measure KSAOs relevant for the aviator job.
- At least some types of psychomotor and information processing tests are likely to exhibit a reasonable level of validity in almost any sample or setting, as evidenced by the Damos (1993) finding of mean uncorrected validities of .18 and .23 for single-task and multiple-task performance-based measures, respectively. This is supported by the range of mean validities from .10 to .32 for measures of dexterity, reaction time, and perceptual speed in Hunter and Burke (1994) and by the Martinussen (1996) mean validity of .20 for psychomotor/information processing tests.
- Measures of spatial ability showed a mean uncorrected validity of .19 in the Hunter and Burke meta-analysis. In the Norwegian Air Force battery, subtests with titles that

appear most like traditional measures of spatial ability (Paper Forming, Rotating Patterns, and Figure Pattern) showed very low validity. However, two other subtests that contain a spatial ability component, Raven's Matrices and Instrument Comprehension, showed much higher validity (mean uncorrected $r = .16$ and .29, respectively). Stage two of the NAF battery also includes spatial ability tests, and they showed very low validity, but this could be due to the extreme restriction of range given that applicants had already been directly screened on spatial abilities during the first stage of the testing process.

- Personality measures, in general, showed low validity for predicting performance in training. However, as noted above, the meta-analytic reviews did not calculate the mean validity for different types of personality traits, and averaging across scales more and less relevant for the aviator job likely obscured the true level of validity that such measures can achieve.

Personality Research in the Aviator Selection Arena

As noted earlier in this report, a great deal of research has been conducted in the area of personality measurement for use in aviator selection, with contradictory results. Lambirth, Dolgin, Rentmeister-Bryant, and Moore (2003) commented, "The US Navy, Air Force, and Army have investigated a variety of personality tests for use in pilot selection batteries. These efforts have had little impact on the selection of pilots or other aircrew because of response bias and the inappropriateness of the clinical measures selected for a homogeneous, non-clinical population. However, personality tests that emphasize positive attributes, rather than psychopathology, and performance-based personality measures, have proven to be more accurate descriptors of personality and predictors of performance" (p. 416).

Job analyses and other studies suggest that non-cognitive characteristics are important for aviator performance. Musson, Sandal, and Helmreich (2004) say, "Superior performance [among pilots] has consistently been linked to a personality profile characterized by a combination of high levels of instrumentality and expressivity along with lower levels of interpersonal aggressiveness. This personality profile has sometimes been referred to as the 'Right Stuff,' suggesting this is the ideal description of an astronaut or pilot. Inferior performance has been linked to personality profiles typified by a hostile and competitive interpersonal orientation . . . (the 'Wrong Stuff') . . . or to low achievement motivation combined with passive-aggressive characteristics ('No Stuff')" (p. 342). The authors point out that these profiles seem to be especially important in terms of working as part of a crew.

As noted above, several of the obstacles to conducting good research in the aviator selection domain are particularly problematic for personality measures. These include the reliance on training outcomes as the criterion measure and summarizing results by averaging validity estimates across several different personality scales.

Findings from General Selection Research Literature

There is a great deal of information about the validity of measurement methods in the general selection research literature. Cognitive ability tests have been shown to predict job performance, particularly technical or "can-do" aspects of job performance, in a wide variety of

jobs (Hunter & Hunter, 1984; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Schmidt & Hunter, 1998; Vernon, 1969). Schmidt and Hunter (1998) collected meta-analytic evidence from a large number of sources and summarized it according to different types of personnel measures, that is, measurement *methods*, for predicting performance in training programs and for predicting overall job performance. Personnel measures with the highest validity for predicting performance in job training programs include general mental ability (GMA) tests (mean r = .56),[3] integrity tests (mean r = .38), peer ratings (mean r = .36), employment interviews (structured and unstructured) (mean r = .35), conscientiousness tests (mean r = .30), and biographical data (mean r = .30). Personnel measures with the highest levels of validity for predicting overall job performance include work sample tests (mean r = .54), GMA tests (mean r = .51), structured employment interviews (mean r = .51), peer ratings (mean r = .49), job knowledge tests (mean r = .48), training and education ratings (mean r = .45), job tryout procedures (mean r = .44), and integrity tests (mean r = .41). As noted previously, there is a high degree of correspondence between *method* and *target* of measurement for some of the personnel measures, for example GMA tests, but a low degree of correspondence for other personnel measures, for example employment interviews.

In many jobs, technical aspects of performance are not the only aspects that matter to the organization. For example, there is a great deal of interest in predicting organizational citizenship (Organ, 1994; Organ & Ryan, 1995), contextual performance (Borman & Motowidlo, 1993), and "will-do" aspects of job performance (Campbell, Hanson, & Oppler, 2001). To identify attributes that underlie these non-technical aspects of job performance, personnel selection researchers turned to the vast literature on non-cognitive attributes. Research in the personality, biodata, and vocational interest domains has clearly shown that measures of non-cognitive attributes can predict job performance (Barrick & Mount, 1991; Gellatly, Paunonen, Meyer, Jackson, & Goffin, 1991; Hunter & Hunter, 1984; McHenry, et al. 1990; Ones, Viswesvaran, & Schmidt, 1993; Tett, Jackson, & Rothstein, 1991), particularly when a careful effort is made to identify and measure attributes that one would expect to underlie different criterion constructs, and when the presence or importance of those criterion constructs is considered for different types of jobs (e.g., Hough, 1992; Hough & Ones, 2002; Hurtz & Donovan, 2000; Mount, Barrick, & Stewart, 1998; Ones & Viswesvaran, 2001a, 2001b; Reilly & Chao, 1982; Robertson & Kinder, 1993). Several researchers have meta-analyzed validity for personality measures, using the "Big 5" model (Norman, 1963; Tupes & Christal, 1961) or some other model (e.g., Hogan, 1991; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990) as an organizing structure. One well-established finding is that measures of conscientiousness appear to be a valid predictor of job performance in virtually all jobs (Barrick and Mount, 1991; Schmidt & Hunter, 1998). The validity of other personality characteristics seems to depend, to a greater extent, on the type of job. For example, extraversion appears to be more valid for predicting performance in sales and managerial jobs than in other types of jobs (Barrick & Mount, 1991; Hough, 1992).

Finally, there is evidence that vocational interests, for example interest in becoming an aviator, can be a valid predictor of relevant job outcomes. It is generally assumed that interest in

---

[3] In the Schmidt and Hunter (1998) meta-analysis, correlations were corrected for criterion unreliability and range restriction (if present).

a particular occupation will lead a person to be motivated to pursue that occupation and motivated to gain knowledge about it. In aviator selection research, there typically has not been a clear distinction between measures of interests and measures of knowledge or background experience, so it is not possible to estimate the likely validity of a stand-alone self-report measure of interest in aviation. There is, however, evidence from the US Army's Project A that scores on a self-report vocational interest inventory are valid for predicting technical job performance in a variety of Army enlisted military occupations (McHenry, et al., 1990; Oppler, McCloy, Peterson, Russell, & Campbell, 2001).

Incremental Predictive Validity

It is clear from the information described in the preceding section that an aviator selection battery focusing on cognitive ability is likely to be a valid predictor of performance in training and on the job. So, is there anything to be gained by including measures of other KSAOs in the aviator selection process? Research suggests that there is. In addition, given the enormous cost of aviator training, even a small increase in validity can offer significant utility to the US Army.

Aviator Selection Research Literature

Most of the research on this topic in the aviator selection arena was conducted by the USAF, and is based on adding the Basic Aviator Test (BAT) to the AFOQT. The BAT consists of several computer-administered tests measuring psychomotor skills, short-term memory, time-sharing ability, and attitudes toward risk-taking. Across several studies, the BAT demonstrated increases in the amount of variance accounted for (i.e., R^2) ranging from zero to .08 (e.g., Carretta, 1987b, 1988; Carretta & Ree, 1996a), with higher incremental validity for criterion measures other than training pass-fail (e.g., Advanced Training Recommendation Board [ATRB] ratings).

Only one study systematically examined the incremental validity of the individual BAT subtests (Carretta & Ree, 1993). This study found that a BAT-Psychomotor composite score and a BAT-Risk composite score *each* (separately), when added to the AFOQT-Aviator composite score, increased the multiple correlation (R) by about .04 for predicting pass/fail and class rank. Adding the BAT measure of Flying Experience to the AFOQT-Aviator composite score increased R by about .07 while adding BAT Information Processing scores did not increase R significantly. Adding all BAT scores to the AFOQT-Aviator composite score increased R by approximately .13 for both pass/fail and class rank (which translates into an increase in amount of variance accounted for, R^2, by about .02).

In research that did not involve the BAT, Ree (2004c) found that two dependent variables derived from the Test of Basic Aviation Skills (TBAS) increased the amount of variance accounted for in basic flight training performance scores by .02 to .03. TBAS includes measures of psychomotor skills, selective attention, spatial ability, and noticing and responding quickly and appropriately to an "emergency." The report does not specify exactly on what the dependent variables are based, but they appear to involve psychomotor and spatial aspects of test performance. Retzlaff, King, and Callister (1995) and Carretta, Retzlaff, and King (1997) report that tests of aviation interest/aptitude included in the AFOQT have been shown to be useful for predicting aviator performance beyond measures of g and of specific cognitive abilities such as verbal, math, spatial, and perceptual speed.

Blower and Dolgin (1990) used a hierarchical regression model to examine the incremental validity exhibited by three tests: (1) Absolute Difference-Horizontal Tracking, (2) Complex Visual Information Processing, and (3) Risk Taking in predicting success in primary flight training over and above that of intelligence and demographic variables. Each resulted in approximately a 3.5 % increase in variance explained. This study also found that a psychomotor/dichotic listening test, a Manikin test (a mental rotation task), and a Baddeley test (an assessment of working memory) did not add incremental validity.

Several other studies used a regression approach to examine the validity of various types of predictor measures for predicting undergraduate training performance (Bartram & Dale, 1985; Carretta, 1989, 1990; Morrison, 1988; Olea & Ree, 1994), but did not report incremental validity. However, the studies did report that psychomotor, spatial orientation, biographical data, working memory, and to a lesser degree personality, all predicted at least some unique variance in undergraduate aviator training.

General Selection Research Literature

In the general selection research literature, Schmidt and Hunter (1998) meta-analytically derived an estimate of the incremental validity likely to occur when any of several personnel measures were added to a measure of general mental ability for predicting a) performance in a training program or b) overall job performance. For predicting performance in a training program, their results show that the greatest incremental validity can be achieved by supplementing a measure of general mental ability with an integrity test or a conscientiousness test (increase in multiple *R* of .11 and .09, respectively). For predicting overall job performance, the greatest incremental validity can be achieved by supplementing a measure of general mental ability with an integrity test (increase in validity of .14), a conscientiousness test (increase in validity of .12), a work sample test (increase in validity of .12), or an employment interview (increase in validity of .09).

Other research has also shown that measures of non-cognitive attributes can provide incremental validity beyond measures of cognitive attributes. This is especially true for predicting non-technical aspects or "will-do" aspects of job performance (Day & Silverman, 1989; Mount, Witt, & Barrick, 2000; Ones & Viswesvaran, 2001b; Oppler, et al., 2001; Robertson & Kinder, 1993; Russell, Mattson, Devlin, & Atwater, 1990; Salgado, 1998) However, incremental validities have also been found in the vocational interest domain, even for cognitive (or "can-do") aspects of job performance (Gellatly, et al, 1991; Hough, Barge, & Kamp, 2001).

Summary of Incremental Validity Evidence

Based on the research evidence described above, there is reason to believe that measures of the following constructs may add incremental validity beyond that achieved by a battery that reliably and accurately measures general intelligence:

- Psychomotor skills
- Working memory
- Aviation interest/knowledge

- Flying experience – although the type of flying experience may make a difference (fixed-wing versus rotary-wing)
- Personality (including factors such as conscientiousness and risk-taking)

Group Differences

As mentioned previously, the Army aviation applicant sample has historically been homogeneous, that is, relatively young, male, and Caucasian, with at least a high school degree and usually with some post-high school education. Some applicants already have or are working on a private pilot's license, but very few are already certified to fly rotary-wing aircraft. Some are already in the military, while others come from the civilian population. In fact, the Army is the only branch of the US military that allows civilians and military enlisted personnel to apply for slots in the aviation training program. (All branches allow military Commissioned Officers to apply for aviator training.) In the Army, those applicants who are accepted, but who are not a Commissioned Officer at the time of application, must complete Warrant Officer training before they enter aviator training.

In the future, it is likely that the applicant population will become more diverse in terms of race and gender but, barring major policy changes, will likely not become more diverse on the other characteristics listed above. One of ARI's objectives is to minimize, to the extent possible, adverse impact exhibited by the new aviator selection battery. The level of adverse impact exhibited by a test battery depends on various factors, including the selection ratio, the general characteristics of the applicant population, and placement of the pass-fail cutoff on a test battery. While the adverse impact cannot be estimated at this point, the research that might help to anticipate how race and gender subgroups are likely to score on an aviator selection test battery can be examined.

Cognitive Ability Tests

Research conducted on cognitive ability tests using military and civilian samples suggests there will be mean score differences on most cognitive ability tests when racial groups are compared, but that the tests will not be unfair to any racial subgroup (Campbell, 1996; Carretta & Ree, 2000; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Russell, Reynolds, & Campbell, 1994; Sackett, Schmitt, Ellingson, & Kabin, 2001; Toquam, Corpe, & Dunnette, 1989; Wise, Welsh, Grafton, Foley, Earles, Sawin, & Divgi, 1992). Research suggests that there will be a standardized mean score difference of 0.6-1.0 between African-Americans and Whites, with Whites scoring higher on average. Other evidence suggests that the Hispanic-White standardized mean score difference will be about half as large as the African American-White subgroup difference, again with the White mean being higher. Finally, Asian subgroups sometimes earn a higher mean score than the White subgroup, and sometimes earn a lower mean score. Many different interpretations of and explanations for these findings have been offered (e.g., educational differences, subtle or overt racism, cultural bias), but no one has yet found a way to entirely explain or eliminate the differences, and efforts to ameliorate the differences have met with limited success (Sackett, et al., 2001; Schmitt, Sackett, & Ellingson, 2002).

Research suggests that there are no gender differences in general cognitive ability (*g*), but that gender differences will appear on specific types of cognitive ability tests. Specifically,

females tend to perform better than males on tests of verbal ability and more poorly than males on tests of spatial, mathematical, and mechanical abilities (Geary, Saults, Liu, & Hoard, 2000; Maccoby & Jacklin, 1974; Maitland, Intrieri, Schaie, & Willis, 2000; Weiss, Kemmler, Deisenhammer, Fleischhacker, & Delazer, 2003; Wise, at al., 1992). Burke (1995) meta-analyzed gender subgroup differences on aviator aptitude tests and reported findings consistent with those from the general literature. As with the race subgroup differences, a variety of explanations have been offered for these findings, for example, differences in socialization experiences, but no one has fully explained or eliminated them to date.

The magnitude of gender differences on spatial ability tests appears to vary considerably with the type of test (Linn & Peterson, 1985), with the largest differences occurring on tests that involve three-dimensional spatial rotation and the smallest differences occurring on tests that involve spatial visualization (e.g., paper-folding tests). Boer (1991) reviewed construct validity evidence for a variety of spatial ability tests and concluded that "the most important aspects of spatial ability are the identification of the optimal solution strategy and, perhaps, a final process called evaluation and confirmation. It seems that the actual execution of the solution process, including mental rotation, is less important."(p. 108).

The US Army's Project A included several different spatial ability tests. Factor analyses suggested that all the tests load on a single underlying factor, but that some of the tests produced much larger race and gender subgroup differences than others (Russell & Peterson, 2001). Specifically, a spatial abilities test called Assembling Objects showed smaller gender differences than other spatial ability tests but was a valid predictor of behavior. A similar pattern of findings, using most of the same spatial ability tests developed during Project A, occurred in a large-scale study focused on revising the ASVAB (Russell, Reynolds, & Campbell, 1994). These researchers recommended adding the Assembling Objects subtest to the ASVAB, a recommendation that has since been enacted.

Psychomotor Tests

Males typically score considerably higher on psychomotor tests than females (Burke, 1995; Carretta, 1997b; McHenry & Rose, 1988; Russell & Peterson, 2001) and the standardized mean score difference is often larger than 1.0. There is much less reported evidence for race subgroup differences on psychomotor tests but Russell and Peterson (2001) found standardized mean score differences ranging from .38 to .87 between African American and White enlisted personnel on Project A psychomotor tests. This finding may be at least partially explained by the correlation between psychomotor and cognitive abilities (Carretta, 1997a; Ree & Carretta, 1992).

Speeded Information Processing Tests

In Project A and in a joint-services project (Russell & Peterson, 2001; Russell, Reynolds, & Campbell, 1994), there were small to no race or gender subgroup differences on speeded measures of information processing, for example, reaction time. Interestingly, males tended to perform somewhat better than females on measures that focus only on perceptual speed, while the reverse was true for measures that focused on both speed and accuracy. Both Carretta (1997b) and Burke (1995) report similar findings when examining gender differences in performance in samples of USAF and UK Royal Air Force aviator applicants respectively.

Personality and Temperament Measures

Research suggests that personality and temperament measures typically show small or no racial subgroup differences (Bobko, Roth, & Potosky, 1999; Hough, 1998; Ones & Viswesvaran, 1998; Russell & Peterson, 2001; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). In contrast, there often are gender differences on personality and temperament inventories and, prior to passage of the Civil Rights Act of 1991, many test batteries used separate within-group norms for scoring and reporting purposes, that is, separate norms for males and females, and for persons of different racial backgrounds. The Civil Rights Act of 1991 prohibits adjusting scores, using different cutoffs, or otherwise altering the results of employment related tests on the basis of race, color, religion, sex, or national origin. As a consequence, the use of within-group norms has essentially disappeared for cognitive ability tests. In the personality measurement arena, it appears that within-group norms are generally accepted when the results will be used for descriptive or diagnostic purposes (e.g., in a counseling setting) but are much more controversial if the results will be used for employment related decisions (see Sackett & Wilk, 1994).

While research suggests there are practically meaningful subgroup differences between males and females on personality or temperament inventories, the direction and size of the difference depends on which personality or temperament characteristic is being measured (Sackett & Wilk, 1994) and, even then, is not always consistent. Furthermore, very little research has been conducted to determine whether or not these differences lead to differential prediction of job performance. Saad and Sackett (2002) analyzed data from the US Army's Project A and found some evidence that personality scores over-predicted female performance, but no evidence of bias.

Sackett and Wilk (1994) reviewed male-female effect sizes on the scales of several well-known personality inventories. The results are difficult to summarize across inventories because, when the inventories were developed, there was no common set of scale labels and no agreed-upon set of underlying constructs. Generally, males scored higher than females on scales measuring dominance, independence, aggression, and risk-taking, while females scored higher than males on scales measuring nurturance, agreeableness, affiliation, and conscientiousness. Many of the differences were not large, however.

The temperament inventory developed as part of Project A (the *Assessment of Background and Life Experience – ABLE)* was developed with the intention of using the same set of norms for both males and females. In a large sample of enlisted US Army Soldiers, the Male-Female effect sizes ranged from .00 to .54 across the 11 ABLE content scales. With the exception of the Physical Condition scale, all of the Male-Female effect sizes were .25 or lower. Female Soldiers scored at least somewhat higher, on average, than males on cooperativeness, conscientiousness, non-delinquency, traditional values, work orientation, internal locus of control, and energy level. Male Soldiers scored at least somewhat higher, on average, than females on emotional stability, self-esteem, dominance, and physical condition (Russell & Peterson, 2001). Finally, Ones and Viswesvaran (1998) meta-analyzed subgroup differences on overt integrity tests and found that females scored .16 standard deviations higher than males.

What Should the Army Measure?

Research focused specifically on aviator selection, as well as general research, clearly suggests that cognitive ability, or general intelligence (*g*), will be an important predictor of aviator performance. Researchers debate the usefulness of identifying more specific abilities within the cognitive ability domain (e.g., Jensen, 1993; Ree & Earles, 1992, 1993; Schmidt & Hunter, 1993; Sternberg & Wagner, 1993), but many personnel selection batteries include measures of different types of cognitive ability, including some combination of:

1. General Reasoning;
2. Spatial Ability;
3. Mechanical Reasoning;
4. Quantitative Ability;
5. Verbal Ability;
6. Multiple-Task Performance (also known as Timesharing or Divided Attention); and,
7. Information Processing (e.g., perceptual speed and accuracy, working memory, cognitive task prioritization).

Research also suggests that including measures of the following abilities and characteristics is also likely to enhance the validity of the overall selection process:

8. Aviation or Helicopter Knowledge
9. Interest in Aviation
10. Flying Experience - although the type of flying experience may make a difference (fixed-wing versus rotary-wing)
11. Normal-Range Personality Characteristics — Based on the aviator selection and general research literature, traits that seem relevant for the aviator job include:
    a. Conscientiousness/Integrity;
    b. Achievement Orientation;
    c. Stress Tolerance/Emotional Stability;
    d. Adaptability/Cognitive Flexibility;
    e. Interpersonal/Crew Interaction skills;
    f. Risk Tolerance;
    g. Internal Locus of Control; and,
    h. Dominance/Potency (including Self-Confidence/Self-Esteem).

These, then, are KSAOs that should be included, at a minimum, in a job analysis study. Some of them may be defined more narrowly than shown here, based on taxonomic work in the

field of individual differences. There are other areas that should be considered in the overall Army aviator selection process, for example, screening for serious medical conditions, neurological deficiencies, or psychological disorders. However, these measures would be designed to "select out" applicants that do not belong in Army aviation training, while the present project is intended to identify those batteries that would be useful in "selecting in" the most qualified applicants.

Review of Existing Aviator Selection Test Batteries

Even using a focused approach to the literature review, more than 150 potentially relevant articles were identified. Rather than rely entirely on a narrative summary, a spreadsheet was developed to summarize standard information about various test batteries and to facilitate comparison of the test batteries when deriving a recommended selection strategy. The following questions were identified as potentially having some bearing on testing recommendations:

1. What subtests, if any, are part of the battery?
2. Who uses the battery?
3. How long does it take to administer the battery?
4. Is the battery already computerized? Web-enabled?
5. Does the battery require non-standard equipment (e.g., joystick, timing card)?
6. What validity evidence is available for the battery?
7. What are the key studies and references describing validation efforts?

An answer for each question was provided (when possible) for a number of current or recently available test batteries, and documented in the aforementioned spreadsheet. The results are shown in Appendix A.

When considering potential measures of non-cognitive characteristics such as conscientiousness, it is clear that no single, existing inventory measures every characteristic that might be important for the aviator job. Therefore, a second spreadsheet was created, shown in Appendix B, to summarize information available for several inventories that have been administered in an aviator selection setting, or that are already owned by the US military. This is not intended to be a comprehensive review of all possible personality inventories. There are dozens of commercially-available personality and biodata inventories that could be used to measure characteristics important for aviators, but none that are specifically designed for aviator selection and none that appear to be significantly more comprehensive or likely to exhibit significantly higher validity than those already available to the US military.

Selection Strategy Recommendations

The following approach was used to develop a recommended selection strategy:

1. Identify KSAOs important for the aviator job through the literature review and a job analysis. In other words, take a construct-oriented approach to this effort.

2. Identify, to the extent possible, existing measures of those KSAOs with known validity.
3. Construct a set of recommendations outlining best-bet choices for predictors that measure critical KSAOs, taking into account what is known about expected subgroup differences.

As noted above, one of the primary considerations in recommending an existing test battery is whether or not there is validity evidence to support its use. Based on the literature review, and as summarized in Appendix A, there are several aviator selection batteries that have demonstrated a reasonable level of validity for predicting Undergraduate Pilot Training, which is typically operationalized as pass/fail, but sometimes also includes measures of training grades or instructor aviator ratings. Almost no one has attempted to predict aviator performance outside of training and, when they have, have not had a great deal of success. This is likely due to issues with criterion quality, as there are significant obstacles to developing good measures of aviator performance on the job.

The most viable candidates for replacing the AFAST appear to be test batteries developed by the US military, several of which are described below.[4] There are also some aviator selection batteries developed by foreign military services or commercial organizations with demonstrated evidence of validity, as shown in Appendix A. The latter test batteries are less viable than batteries developed by the US military because: 1) there is no evidence that they are any more valid than test batteries developed by the US military, and 2) it would likely be difficult and/or expensive for the US Army to gain access to them.

The recommendations resulting from this review are presented in overview form in Appendix C. It would seem to be an efficient use of Army testing resources to create a two-stage testing process. The first stage would include measures of cognitive abilities such as spatial ability, mechanical reasoning, verbal ability, numerical reasoning, and perceptual speed and accuracy, as well as a measure that would attempt to tap motivation to become an aviator, such as an Information subtest. The Army may be able to take advantage of the fact that the US Navy has a web-enabled aviator selection battery that currently consists of a reasonable set of cognitive tests and an Information subtest that assesses aviation and nautical knowledge. Including a non-cognitive inventory in Stage 1 is also recommended. Such an inventory may provide incremental validity beyond the cognitive test battery, and it may help ameliorate race subgroup differences on the cognitive tests. The inventory could include scales from several different inventories that have been developed by the US military. The Stage 1 battery could be administered via the Internet on any standard desktop computer and would not require any non-standard peripherals or hardware. Thus, it could be administered virtually anywhere that a computer is available, along with a reliable Internet connection and a test control officer.

---

[4] Re-using any of the existing AFAST subtests was not considered because Army researchers believe that the content may have been compromised over the several years in which it has been used.

The second stage of the test battery would focus on psychomotor skills and multiple-task performance. These types of tests are often combined and labeled "performance-based" measures. Alternately, given practical considerations, this Stage 2 test battery might assist in the classification of selected Army aviators into mission/aircraft types.

Best Bet Predictor Measures

Stage 1: Cognitive Measures

Aviator Selection Test Battery (ASTB). The ASTB is the US Navy's primary aviator selection instrument. It grew out of the Pensacola 1000 Pilot Study, which examined over 60 psychological, psychomotor, and physical tests (North & Griffin, 1977). The current version of the ASTB includes subtests measuring Reading Comprehension, Mathematical Ability, Mechanical Comprehension, Spatial Apperception, and Aviation and Nautical Interests. Navy researchers are currently building an adaptive version of the ASTB and anticipate transitioning to adaptive testing within three to five years.

The ASTB subtests are used to create several composite scores, including the Academic Qualification Rating (AQR) and the Pilot Flight Aptitude Rating (PFAR). Validity data for FY98-FY04 are summarized or reported in graphic form in a series of ASTB Workshop Briefing Slides (Operational Psychology Department, 20 July 2004). Navy researchers found that AQR scores predict performance in ground school (r = .46 for USN student aviators and r = .39 for USMC student aviators) while PFAR scores predict performance in primary flight school (Primary NSS) (r = .32 USN student aviators and r = .21 for USMC student aviators). This research suggest reported that validity for predicting attrition from aviator training was in the high teens for student aviators in both the USN and the USMC, using AQR or PFAR scores. Sample sizes were not provided in the briefing slides, but include thousands of cases (personal communication, Captain John Schmidt, Operational Psychology Department, USN, October 29, 2004).

The US Navy has developed a web-administration system for the ASTB, called Automated Pilot Examination (APEX). The APEX system is being widely used throughout the Navy and is expected to account for the bulk of ASTB administrations by the end of FY 2005 (personal communication, Captain John Schmidt, USN, October 29, 2004). This is currently the only web-administered aviator selection test battery.

The ASTB was designed to select Commissioned Officers who will enter training to become a Navy or USMC aviator. All Commissioned Officers, by definition, have completed a four-year college degree. Therefore, to ensure that the ASTB is not too difficult for the Army aviator applicant population (which includes persons with less than a four-year college degree), the US Navy administered the ASTB to a sample of incoming Army student aviators. In February 2005, the Operational Psychology Department of the Naval Operational Medicine Institute administered a paper-and-pencil version of the ASTB to 73 student aviators at Ft. Rucker, AL. The Navy scored the data and provided summary information to the ARI monitor and PDRI project team. There was a reasonable degree of variability in the Army scores, and no evidence that it was too difficult for Army student aviators (i.e., no floor effect). The Navy also provided, for comparison purposes, ASTB summary data for Navy personnel with varying levels of education. (The ASTB is administered to Navy personnel for some purposes other than

aviator selection.) Overall, the Army sample scores were similar to those in a mixed-education Navy sample, and to those in a sample of Navy personnel who had at least a bachelor's degree. The Army aviator sample included in this effort had already been selected via the AFAST, and thus does not represent the full range of intelligence in the Army aviator applicant population. Nevertheless, it appears that the ASTB will not prove to be too difficult, overall, for the US Army aviator applicant sample.

Air Force Officer Qualification Test (AFOQT). The US Air Force developed the AFOQT in the early 1950s as a tool for selecting civilian applicants for officer precommissioning training programs and for classifying commissionees into aircrew job specialties (Rogers, Roach, & Short, 1986; Skinner & Ree, 1987). The Air Force has periodically revised the AFOQT to update items, ensure test security, and improve predictive validity. The first form of the AFOQT was implemented in 1953, Form R is currently in use, and Form S is scheduled for implementation in the near future. Form R has 16 subtests and Form S has 11 subtests that tap verbal, quantitative, spatial, and mechanical aptitudes. Form S also includes a measure of non-cognitive characteristics, called the Self-Description Inventory. Scores on the AFOQT subtests are used to form five distinct but partially overlapping composites: Pilot, Navigator-Technical, Academic Aptitude, Verbal, and Quantitative (Sperl & Ree, 1990). The Pilot and Navigator-Technical composites are used for classification into Undergraduate Pilot Training (UPT) and Undergraduate Navigator Training (UNT), respectively. The AFOQT has been validated for more than 36 officer jobs as a predictor of technical training grades (Arth, 1986; Carretta & Ree, 1998). Carretta and Ree (1994) found that the AFOQT showed a multiple correlation of .20 for predicting rank in UPT, and Shore and Gould (2003) reported a multiple correlation (uncorrected) of .34 with UPT final grade for Form S of the AFOQT. According to Carretta (2002), "the predictiveness of the AFOQT for aviator training performance comes almost entirely from its measurement of *g* and aviation job knowledge."(p. 1).

As new forms of the AFOQT have been constructed in recent years, key features of the subtests have deliberately been held constant to ensure equivalent measurement. Thus, the more recent versions are equivalent in terms of subtest content, subtest length, item difficulty, testing time, and stylistic features. Further, about one-half of the items in each form are taken directly from the previous form, and analyses are conducted to equate the new form to the old (Glomb & Earles, 1997). None of the AFOQT forms, to date, have been computerized, but the USAF intends to develop this capability.

Like the ASTB, the AFOQT is designed primarily for use with a Commissioned Officer population. Therefore, USAF provided access to, and permission to analyze, a normative database containing scores on the soon-to-be-implemented AFOQT Form S for a Basic Military Training (BMT) enlisted personnel likely to apply for the Airman Education and Commissioning Program (n = 509), Air Force Reserve Officer Training Cadets (n = 679), and Officer Training School cadets (n = 462). The analyses are described in Gould and Damos (2005). They conclude, "As expected, the AFOQT was more difficult for the Air Force enlisted personnel than for other commissioning source applicants. However, the subtest and composite score distributions are sufficient to discriminate well between enlisted personnel if the AFOQT or similar aptitude test is used for [aviator] selection" (p. 1). If the US Army chooses to implement the AFOQT, these authors recommend that Army-specific norms and passing score(s) be established.

Cognitive Prioritization (Popcorn Test). The cognitive prioritization test follows a format originally developed by NASA researchers, and is colloquially known as a "popcorn" test. It is a measure of cognitive processing, specifically the ability to prioritize several moving stimuli that appear on a computer screen. No operational pilot selection test battery has included a popcorn test, although other test batteries, for example, Wombat© (Aero Innovation, 1998), likely measure the same or a similar underlying ability. It is recommended that the US Army include this type of test in its Army aviator selection battery because this ability may become increasingly important in the future, as the cognitive load associated with flying rotary-wing aircraft increases. Scores on this test may also be related to measures of situational awareness.

Perceptual Speed and Accuracy. One possible measure of perceptual speed and accuracy is the Table Reading test that is a subtest of the AFOQT. This test has been in use for aviator selection since 1942. It continues to account for unique variance in prediction of aviator performance, and is part of the AFOQT Pilot Composite score. A commercial version of the test is also available.

Alternatively, it would be possible to develop a new measure of perceptual speed and accuracy, using stimulus materials that are face valid for Army aviators. PDRI has developed many different measures of perceptual speed and accuracy, and could do so efficiently in the current project.

Stage 1: Non-Cognitive Measures

Test of Adaptable Personality (TAP). The TAP was developed by the US Army for use in training and developing Special Forces Soldiers and officers. It consists of biodata items that were written to target constructs such as achievement orientation, fitness motivation, cognitive flexibility, peer leadership, and interpersonal skills. In Special Forces samples, the achievement orientation, fitness motivation, and cognitive flexibility scales have proven valid for predicting peer and supervisor ratings of performance (personal communication, R. Kilcullen, November, 2005; Kilcullen, Goodwin, Chen, Wisecarver, & Sanders, undated; Kilcullen, Mael, Goodwin, & Zazanis, 1999).

Assessment of Individual Motivation (AIM). The AIM is a forced-choice non-cognitive inventory that measures several constructs potentially important for aviator selection. It was developed by researchers at ARI, and was developed to measure most of the same constructs as the Assessment of Background and Life Experiences (ABLE) developed during Project A. In Project A, the ABLE was predictive of volitional aspects of performance in a variety of military enlisted jobs, and it exhibited incremental validity when added to a cognitive test battery (Russell & Peterson, 2001). However, the ABLE was never implemented for selection purposes due to concerns about its fakability (White, Young, & Rumsey, 2001).

The AIM specifically addresses fakability concerns by using the forced-choice methodology. This methodology has long been suggested as a way to make an inventory resistant to faking, and there is evidence to support this claim, some of it specifically based on the AIM (Jackson, Wroblewski, & Ashton, 2000; White, et al., 2001). ARI is also currently funding efforts to explore an Item Response Theory (IRT)-based approach to administering and

scoring the AIM, in an attempt to make it even more resistant to faking (Stark, Chernyshenko, & Drasgow, 2003).

To date, research on the AIM has focused primarily on predicting attrition, but there is some evidence that it predicts job performance and personal discipline among correctional officers in military prisons as well as success in explosive ordinance disposal training for military personnel (White, et al., 2001). Project A results suggest it is reasonable to believe that the AIM will predict volitional aspects of job performance for the Army aviator job, because it measures characteristics important for performing that job.

The AIM is currently used for operational recruit screening as part of the US Army's GED Plus program, and it has shown promise for use in pre-enlistment screening of Non High School Graduate (NHSG) recruits (White, Young, Heggestad, Stark, Drasgow, & Piskator, 2004, 2005). It is also being evaluated for potential use in screening of US Army recruiters and drill sergeants. Researchers have also developed various scoring methods in an effort to enhance the validity of the AIM for predicting attrition, including empirical scoring procedures (White, Young, Heggestad, Stark, Drasgow, & Piskator, 2005), an IRT-based scoring approach (Chernyshenkso, Stark, & Drasgow, 2003), and a decision tree approach (Lee & Drasgow, 2003).

Self-Description Inventory Plus (SDI+). The SDI was developed by the USAF, and is currently considered an experimental subtest within Form S of the AFOQT. It was originally developed to measure the Big Five personality factors. In recent years, USAF researchers wrote two additional scales to measure Team Orientation and Commitment to Military Service (Service Orientation). It contains 220 items (see Christal, et al., 1997). According to USAF researchers, the value of the SDI is in generating profiles for people and ultimately profiles for organizations and job families to facilitate person-job match and strategic force development (J. Weissmuller, personal communication, February 28, 2005). It is not specifically intended for personnel selection. Nevertheless, validity data are currently being collected in a broad USAF sample, including some aviators.

Armstrong Laboratory Aviation Personality Scale (ALAPS). The ALAPS was also developed by the USAF (Retzlaff, King, Callister, Orme, & Marsh, 2002). It includes five "personality" scales (confidence, socialness, aggressiveness, orderliness, and negativity), six "crew interaction" scales (dogmatism, deference, team orientation, organization, impulsivity, and risk-taking), and four "psychopathology" scales (affective lability, anxiety, depression, and alcohol abuse). A large-scale validation study is currently underway by the USAF. The US Navy is also planning to conduct validation research on this inventory. However, further investigation into this inventory revealed the unfortunate fact that the items and scoring key have been published in a USAF technical report that is available to members of the general public who are savvy enough to locate it. Therefore, it would be unwise for the US Army to use this inventory for selection.

New Non-cognitive Scales. Based on our review of the existing inventories, there are several non-cognitive characteristics that may be predictive of aviator performance that are not measured by any of the inventories readily accessible to the US Army. Therefore, it may be

advisable to write new scales targeting these characteristics, emulating the style and format of the items in the TAP.

Stage 2: Psychomotor Skills and Multiple-Task Performance (Performance-Based Measures)

Test of Basic Aviation Skills (TBAS). This test battery was developed by the USAF as a replacement for the BAT. It includes three subtests designed to measure spatial orientation (tracking tasks) and multiple task performance skills (tracking plus directed listening), as well as the ability to make decisions under stress. TBAS is scheduled for fielding in 2006 and the US Navy is also considering adding it to their aviator selection process. Ree (2003) analyzed TBAS data from USAF aviator trainees (n = 531) who had already been selected on other measures and found that the spatial orientation and decision-making subtests showed low but significant correlations with various training criterion measures. The multiple correlation for predicting a combined training performance measure (based on check ride scores, instructor ratings, and quiz scores, among other things) was .33; the multiple correlation for predicting UPT pass/fail was .31. It does not appear that these correlations were corrected for shrinkage but the author notes that they were downwardly biased due to a high degree of range restriction on the predictor measures. There are some concerns about the stability of scores on the decision-making subtest. Ree (2004b) examined 90-day and 180-day test-retest reliability in a small sample (n = 126) of USAF aviator trainees. Reliability was very low for the decision-making subtest (.15) and acceptable for the other subtest scores (.56-.75). Further investigation of TBAS with USAF personnel revealed the unfortunate fact that no documentation regarding the computer programming appears to exist, nor any documentation about how the dependent variables are calculated. For this reason, it is not recommended that the TBAS be used for Army aviator selection unless and until program documentation can be located.

Wombat©. The Wombat© (Aero Innovation, 1998) is a commercially-available, computerized test battery that involves learning and operating a complex system. It does not involve discrete subtests, but rather involves continuous performance on a primary tracking task, with secondary performance on any of three "bonus" tasks. The bonus tasks are worth varying amounts of points at different times. All of the measures are combined to create a total efficiency score. During the testing period, examinees are given continuous feedback on their performance which can help them maximize their task performance strategies, to the extent that they have the attentional and cognitive capacity to do so. There has been little published on the validity of the Wombat©, but two studies suggest that scores are correlated with academic performance in flight school and flight hours (Cain, 2002; Frey, Thomas, Walton, & Wheeler, 2001). The Wombat© has been used extensively for aviator selection in Canada, but has not been used operationally by the US military, as the advertised pricing is prohibitive.

New Performance-Based Measure. If neither the TBAS nor the Wombat© are viable alternatives, it is recommended that the US Army develop its own performance-based measure of psychomotor skills and multiple-task performance. This recommendation is being made because there does not appear to be another performance-based measure that: 1) has proven validity; 2) is programmed in a modern programming language; and, 3) is readily available and free to the US Army.

The new test battery could include subtests similar to psychomotor tests with a long history and proven validity, for example, the Complex Coordination test and the Rotary Pursuit test, but programmed in a modern programming language. Multiple-task performance could be assessed by combining a directed listening test, or some other secondary task, with a psychomotor task. For example, it might be possible to use the TBAS as a model for development but with careful documentation of the programming and development of scoring variables.

Conclusions

This report presents a review of a great deal of research in the aviator selection and general personnel selection domains. That information was used to identify KSAOs that should be included in a job analysis study focusing on the Army aviator job. It was further used to develop a recommended strategy for an Army aviator selection battery.

Research focused specifically on aviator selection, as well as general personnel selection research, clearly suggests that cognitive ability, or general intelligence (*g*), will be an important predictor of aviator performance. More specific cognitive abilities that may be of importance include: general reasoning; spatial ability; mechanical reasoning; quantitative ability; verbal ability; multiple-task performance (also known as timesharing or divided attention); and information processing (e.g., perceptual speed and accuracy, working memory, cognitive task prioritization). Research also suggests that measures of aviation or helicopter knowledge, interest in aviation, flying experience, and normal-range personality characteristics are likely to enhance the validity of the overall selection process. Non-cognitive traits that seem relevant for the aviator job include: conscientiousness/integrity; achievement orientation; stress tolerance/emotional stability; adaptability/cognitive flexibility; interpersonal/crew interaction skills; risk tolerance; internal locus of control; and, dominance/potency (including self-confidence/self-esteem).

The results of this review, then, suggest a selection strategy for Army aviation that includes measures of cognitive abilities such as spatial ability, mechanical reasoning, verbal ability, numerical reasoning, perceptual speed and accuracy, and cognitive prioritization, as well as a measure that would attempt to tap motivation to become an aviator. In addition, incremental validity may be achieved by including non-cognitive measures such as the TAP and AIM, as well as other normal-range personality inventories.

From a practical perspective, the test battery could be administered via the Internet on any standard desktop computer and would not require any non-standard peripherals or hardware. Thus, it could be administered virtually anywhere that a computer is available, along with a reliable Internet connection and a test control officer. In fact, the Army may be able to take advantage of the fact that the US Navy has a web-enabled aviator selection battery that currently consists of a reasonable set of cognitive tests.

The addition of measures that focus on psychomotor skills and multiple-task performance, often labeled "performance-based" measures, is recommended. However, practical constraints on time and resources might suggest that these tests be considered as candidates for inclusion in an aviator tracking battery, to assist in the classification of selected Army aviators into mission/aircraft types. This recommended "Stage 2" in the selection/classification process

is, in fact, the next scheduled research and development effort for the ARI Rotary-Wing Aviation Research Unit at Fort Rucker.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*, 3-27.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*, 288-318.

Ackerman, P. L. (1990). A correlational analysis of skill specificity: Learning, abilities, and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 883-901.

Ackerman, P. L., Kanfer, R., & Goff, M. (1995). Cognitive and non-cognitive determinants and consequences of complex skill acquisition. *Journal of Experimental Psychology: Applied, 1*, 270-304.

Ackerman, P. L., & Woltz, D. J. (1994). Determinants of learning performance in an associative memory/substitution task: Task constraints, individual differences, volition, and motivation. *Journal of Educational Psychology, 86*, 487-515.

*Aero Innovation Inc. (1998, June). *WOMBAT-CS Candidate's Manual (21st edition) (for software version CS 4.9)*. Montreal, Quebec: Author.

Ambler, R. K., & Smith, M. J. (1974). *Differentiating aptitude factors among current aviation specialties* (NAMRL-1207). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

*Anesgart, M. N., & Callister, J. D. (1999). Predicting training success with the NEO: The use of logistic regression to determine the odds of completing a pilot's screening program. From *Proceedings of the Tenth International Symposium on Aviation Psychology,* Ohio State University: Department of Aerospace Engineering.

Arth, T. O. (1986). *Validation of the AFOQT for non-rated officers* (AFHRL-TP-85-50). Brooks AFB, TX: Air Force Human Resources Laboratory, Air Force Systems Command.

*Arth, T. O., Steuck, K. W., Sorrentino, C. T., & Burke, E. F. (1990). *Air Force Officer Qualifying Test (AFOQT): Predictors of Undergraduate Pilot Training and Undergraduate Navigator training success* (Interim Technical AFHRL-TP-89-52): Air Force Human Resources Laboratory.

*References marked with an asterisk are cited in one of the appendices.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

*Bartram, D. (1987). The development of an automated testing system for aviator selection: The MICROPAT project. *Applied Psychology: An International Review, 36*(3/4), 279-298.

*Bartram, D., & Dale, H. C. A. (1982). The Eysenck Personality Inventory as a selection test for military pilots. *Journal of Occupational Psychology, 55*, 287-296.

Bartram, D., & Dale, H. C. A. (1985). The prediction of success in helicopter pilot training. *Report of the XVI Conference of the Western European Association of Aviation-Psychology* (pp. 92-101). Helsinki, Finland: Finnair Training Center.

*Bishop, S. L., Faulk, D., & Santy, P. A. (1996). The use of IQ assessment in astronaut screening and evaluation. *Aviation, Space and Environmental Medicine, 67*(12), 1130-1138.

*Blower, D. J. (1998). *Psychometric equivalency issues for the APEX system* (NAMRL Special Report 98-1). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

*Blower, D. J., & Dolgin, D. L. (1990). An evaluation of performance based tests designed to predict success in primary flight training. Paper presented at the 34th Annual Meeting of the Human Factors Society.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*, 561-589.

Boer, L. C. (1991). Spatial ability and the orientation of pilots. In R. Gal, & A. D. Mangelsdorf (Eds.), *Handbook of Military Psychology*. New York: Wiley & Sons.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt, & W. C. Borman (Eds.), *Personnel selection in organizations*. San Francisco, CA: Jossey-Bass Publishers.

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. *International Journal of Selection and Assessment, 9*, 52-69.

Burke, E. (1995). Male-female differences on aviation selection tests: Their implications for research and practice. In N. Johnston, R. Fuller, & N. McDonald (Eds.), *Aviation Psychology: Training and Selection* (pp. 188-193). Aldershot, England: Avebury Aviation.

*Burke, E., Hobson, C., & Linsky, C. (1997). Large sample validations of three general predictors of pilot training success. *The International Journal of Aviation Psychology, 7*, 225-234.

*Burke, E., Kitching, A., & Valsler, C. (1997). The Pilot Aptitude Tester (PILAPT): On the development and validation of a new computer-based test battery for selection of pilots. In R. S. Jensen, & L. Rakovan (Eds.), *Ninth International Symposium on Aviation Psychology* (pp. 1286-1291). The Ohio State University: Department of Aerospace Engineering, The Ohio State University.

*Cain, R. E. (2002). *The relationships of metacognition, self-efficacy, and educational and/or flight experience to situational awareness in aviation students.* Unpublished dissertation, University of Missouri, Columbia, MI. [Dissertation Abstracts International Section A: 2986.]

*Caldwell, J. A., O'Hara, C., Caldwell, J. L., Stephens, R. L., & Krueger, G. P. (1993). Personality profiles of U.S. Army helicopter pilots screened for special operations duty. *Military Psychology, 5*, 187-199.

*Callister, J. D., King, R. E., & Retzlaff, P. (1996). Cognitive assessment of USAF pilot training candidates. *Aviation, Space and Environmental Medicine, 67*(12), 1124-1129.

*Callister, J. D., King, R. E., Retzlaff, P. D., & Marsh, R. W. (1999). Revised NEO personality inventory profiles of male and female U.S. Air Force pilots. *Military Medicine, 164*(12), 885-890.

Campbell, J. P. (1996). Group differences and personnel decisions: Validity, fairness, and affirmative action. *Journal of Vocational Behavior, 49*, 122-158.

Campbell, J. P., Hanson, M. A., & Oppler, S. H. (2001). Modeling performance in a population of jobs. In J. P. Campbell, & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification.* Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, J. P., Harris, J. H., & Knapp, D. J. (2001). The Army Selection and Classification Research Program: Goals, overall design, and organization. In J. P. Campbell, & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification.* Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, J. P., & Knapp, D. J. (Eds.) (2001). *Exploring the limits in personnel selection and classification.* Mahwah, NJ: Lawrence Erlbaum Associates.

*Carretta, T. R. (1987a). *Basic attributes tests (BAT) system: Development of an automated test battery for pilot selection* (AFHRL-TR-87-9). Brooks AFB, TX: Air Force Human Resources Laboratory.

*Carretta, T. R. (1987b). *Field dependence independence and its relationship to flight training performance* (Interim Technical Paper AFHRL-TP-87-36). San Antonio, TX: Brooks AFB, Air Force Human Resources Laboratory.

*Carretta, T. R. (1987c). *Spatial ability as a predictor of flight training performance* (Interim Technical AFHRL-TP-86-70). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

*Carretta, T. R. (1987d). *Time-sharing ability as a predictor of flight training performance* (Interim Technical AFHRL-TP-86-69). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

*Carretta, T. R. (1988). *Relationship of encoding speed and memory tests to flight training performance* (AFHRL-TP-87-49). San Antonio, TX: Brooks AFB, Air Force Human Resources Laboratory.

Carretta, T. R. (1989). USAF pilot selection and classification systems. *Aviation, Space and Environmental Medicine, 60*, 46-49.

Carretta, T. R. (1990*). Basic attributes test (BAT): A preliminary comparison between Reserve Officer Training Corps (ROTC) and Officer Training School (OTS) pilot candidates* (AFHRL-TR-89-50). Brooks AFB, TX: Air Force Human Resources Laboratory.

Carretta, T. R. (1997a). Group differences on US Air Force pilot selection tests. *International Journal of Selection and Assessment, 5*, 115-127.

Carretta, T. R. (1997b). Sex differences on US Air Force pilot selection tests. *Proceedings of the Ninth International Symposium on Aviation Psychology* (pp. 1292-1297). Columbus, OH: The Ohio State University.

*Carretta, T. R. (2000). US Air Force pilot selection and training methods. *Aviation, Space and Environmental Medicine, 71*, 950-956.

Carretta, T. R. (2002). Common military pilot selection practices. *Human Systems IAC Gateway, XIII(1),* 1-4.

*Carretta, T. R., & Ree, M. J. (1993). *Pilot Candidate Selection Method (PCSM): What makes it work?* (AL-TP-1992-0063). Brooks AFB, TX: Manpower and Personnel Research Division, Human Resources Directorate, Air Force Systems Command.

*Carretta, T. R., & Ree, M. J. (1994). Pilot-candidate selection method: Source of validity. *International Journal of Aviation Psychology, 4*, 103-118.

Carretta, T. R., & Ree, M. J. (1996a). Central role of g in military pilot selection. *The International Journal of Aviation Psychology, 6*(2), 111-123.

Carretta, T. R., & Ree, M. J. (1997a). Negligible sex differences in the relation of cognitive and psychomotor abilities. *Personality and Individual Differences, 22*, 165-172.

Carretta, T. R., & Ree, M. J. (1997b). A preliminary evaluation of causal models of male and female acquisition of pilot skills. *The International Journal of Aviation Psychology, 7*(4), 353-364.

Carretta, T. R., & Ree, M. J. (1998). *Factor structure of the Air Force Officer Qualifying Test: Analysis and comparison* (AL/HR-TP-1997-0005). Mesa, AZ: Air Force Materiel Command, Air Force Research Laboratory, Human Resources Directorate, Aircrew Training Research Division.

Carretta, T. R., & Ree, M. J. (2000). *Pilot selection methods* (AFRL-HE-WP-TR-2000-0116). Wright-Patterson AFB, OH: Human Effectiveness Directorate, Crew System Interface Division.

Carretta, T. R., & Ree, M. J. (2003). Pilot selection methods. In P. S. Tsang, & M. A. Vidulich (Eds.), *Principles and practice of aviation psychology* (pp. 357-396). Mahwah, NJ: Lawrence Erlbaum Associates.

*Carretta, T. R., Ree, M. J., & Callister, J. D. (1999). *Factor structure of the CogScreen-Aeronautical Edition Test Battery* (AFRL-HE-AZ-TR-1998-0076): Mesa, AZ: Air Force Materiel Command, Air Force Research Laboratory, Human Resources Directorate, Aircrew Training Research Division.

Carretta, T. R., Retzlaff, P. D., & King, R. E. (1997). *A tale of two test batteries: A comparison of the Air Force Officer Qualifying Test and the Multidimensional Aptitude Battery* (AL/HR-TP-1997-0052). Brooks AFB, TX: Air Force Materiel Command, Air Force Research Laboratory, Human Resources Directorate, Aircrew Training Research Division.

*Carretta, T. R., Zelenski, W. E., & Ree, M. J. (1997). *Basic Attributes Test retest performance* (AL/HR-TP-1997-0040). Mesa, AZ: Air Force Materiel Command, Air Force Research Laboratory, Human Resources Directorate, Aircrew Training Research Division.

Chernyshenko, O. S., Stark, S. E., & Drasgow, F. (2003). Predicting attrition of Army recruits using optimal appropriateness measurement. In *Proceedings of the 45th Annual Conference of the Military Testing Association* (pp. 317-322), Pensacola, FL.

*Chidester, T. R., & Foushee, H. C. (1991). Leader personality and crew effectiveness: A full-mission simulation experiment. In R. S. Jensen (Ed.), *Proceedings of the Fifth international symposium on aviation psychology, Vol.II.* Columbus, OH: The Ohio State University.

Christal, R. L. (1975). Personality factors in selection and flight proficiency. *Aviation, Space and Environmental Medicine, 46*, 309-311.

*Christal, R., Barucky, J. M., Driskill, W. E, & Collis, J. M. (1997). *The Air Force Self Description Inventory (AFSDI): A summary of continuing research* (Informal Technical Final Report F33615-91-D-0010). San Antonio, TX: Metrica, Inc.

Damos, D. L. (1993). Using meta-analysis to compare the predictive validity of single- and multiple-task measures to flight performance. *Human Factors, 35*(4), 615-628.

Damos, D. L. (1996). Aviator selection batteries: Shortcomings and perspectives. *The International Journal of Aviation Psychology, 6*(2), 199-209.

*Davis, W., Koonce, J., Herold, D., Fedor, D., & Parsons, C. (1997). Personality variables and simulator performance in the prediction of flight training performance, *Proceedings of the Ninth International Symposium of Aviation Psychology* (pp. 1105-1109). Columbus, OH: The Ohio State University.

Day, D. V., & Silverman, S. B. (1989). Personality and job performance: Evidence of incremental validity. *Personnel Psychology, 42*, 25-36.

*Delaney, H. D. (1990). *Validation of dichotic listening and psychomotor task performance as predictors of primary flight training criteria: Highlighting relevant statistical issues* (Technical report NAMRL-1357). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Department of the Army. (1981). *Army Regulation 611-85, Aviation Warrant Officer Training.* Washington, D.C.: Headquarters.

Department of the Army. (1994). *Army Regulation 135-100, Appointment of Commissioned and Warrant Officers of the Army.* Washington, D.C.: Headquarters.

Department of the Army. (1996). *Army Pamphlet 600-11, Warrant Officer Professional Development.* Washington, D.C.: Headquarters.

Department of the Army. (1999). *Army Circular 601-99-1, Warrant Officer Procurement Program.* Washington, D.C.: Headquarters.

Department of the Army. (1999). *Army Memo 600-2, Policies and Procedures for Active-Duty List Officer Selection Boards.* Washington, D.C.: Headquarters.

Department of the Army. (1999). *Army Regulation 135-210, Order to Active Duties as Individuals Other than a Presidential Selected Reserve Call-up, Partial or Full Mobilization.* Washington, D.C.: Headquarters.

Department of the Army. (2002). *Army Regulation 611-5, Army Personnel Selection and Classification Testing.* Washington, D.C.: Headquarters.

Department of the Army. (2003). *Army Regulation 611-110, Selection and Training of Army Aviation Officers (Revised).* Washington, D.C.: Headquarters.

Dolgin, D. L., & Gibb, G. D. (1988). *A review of personality measurement in aircrew selection* (NAMRL-Monograph-36). Pensacola, FL: Naval Aeromedical Research Laboratory.

*Duke, A. P., & Ree, M. J. (1996). Better candidates fly fewer training hours: Another time testing pays off. *International Journal of Selection and Assessment, 4*(3), 115-121.

Fleishman, E. A. (1967). Performance assessment based on an empirically derived task taxonomy. *Human Factors, 9*, 349-366.

Fleishman, E. A. (1972). On the relation between abilities, learning, and human performance. *American Psychologist, 27*, 1017-1032.

Fleishman, E. A., & Hempel, W. E. (1954). Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika, 19*, 239-252.

Fleishman, E. A., & Mumford, M. D. (1988). Ability requirements scales. In S. Gael (Ed.), *Job analysis handbook for business, industry, and government* (Vol. 2, pp. 917-935). New York, NY : Wiley.

Foushee, H. C., & Helmreich, R. L. (1986). Group Interaction and flightcrew performance. In E. L. Wiener, & D. C. Nagel (Eds.). *Human Factors in Modern Aviation*. New York, NY: Academic Press.

Franzen, R., & McFarland, R. A. (1945). *Detailed statistical analysis of data obtained in the Pensacola study of naval pilots* (Report 41). Washington, DC: Civil Aeronautics Administration.

*Frey, B. F., Thomas, M., Walton, A. J., & Wheeler, A. (2001). *WOMBAT as an example of situational awareness testing in pilot selection: An argument for the alignment of selection training, and performance.* Paper presented at the 11th International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

*Fry, G. E., & Reinhardt, R. F. (1969). Personality characteristics of jet pilots as measured by the Edwards Personal Preference Schedule. *Aerospace Medicine, 40*, 484-486.

*Garvin, J. D., Acosta, S. C., & Murphy, T. E., II (1995). Flight training selection using simulators — a validity assessment. In R.S. Jensen (Ed.), *Proceedings of the Eighth International Symposium on Aviation Psychology* (pp. 1132-1136). Columbus, OH: The Ohio State University.

Geary, D. C., Saults, S. J., Liu, F., & Hoard, M. K. (2000). Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology, 77*, 337-353.

Geist, C. R., & Boyd, S. T. (1980). Personality characteristics of Army helicopter pilots. *Perceptual and Motor Skills, 51*(1), 253-254.

Gellatly, I. R., Paunonen, S. V., Meyer, J. P., Jackson, D. N., & Goffin, R. D. (1991). Personality, vocational interest, and cognitive predictors of managerial job performance and satisfaction. *Personality and Individual Differences, 12*, 221-231.

*Glomb, T. M., & Earles, J. A. (1997). *Air Force qualifying test (AFOQT): Forms Q development, preliminary equating and operational equating* (AL/HT-RP-1996-0036). Air Force Materiel Command, Air Force Research Laboratory, Human Resources Directorate, Manpower and Personnel Research Division.

*Gopher, D. (1982). A selective attention test as a predictor of success in flight training. *Human Factors, 24*(2), 173-183.

*Gopher, D., & Kahneman, D. (1971). Individual differences in attention and the prediction of flight criteria. *Perceptual and Motor Skills, 33*, 1335-1342.

*Gordon, H. W., & Leighty, R. (1988). Importance of specialized cognitive function in the selection of military pilots. *Journal of Applied Psychology, 73*(1), 38-45.

*Gough, H. G., & Bradley, P. (1996). *CPI Manual* (3rd Ed.). Mountain View, CA: Consulting Psychologists Press.

Gould, R. B., & Damos, D. L. (2005). *Feasibility of developing a common US Army Helicopter Pilot Candidate Selection System: Analysis of US Air Force data.* Gurnee, IL: Damos Aviation Services.

*Gould, R. B., & Shore, C. W. (2002). *Reduction of AFOQT Administration Time.* San Antonio, TX: Operational Technologies Corporation.

*Gregorich, S., Helmreich, R. L., Wilhelm, J. A., & Chidester, T. (1989). Personality based clusters as predictors of aviator attitudes and performance. In R.S. Jensen (Ed.), *Proceedings of the 5th International Symposium on Aviation Psychology* (pp. 686-691). Columbus, OH: The Ohio State University.

Griffin, G. R., & Koonce, M. J. (1996). Review of psychomotor skills in pilot selection research of the U.S. military services. *International Journal of Aviation Psychology, 6*(2), 125-147.

Griffin, G. R., & McBride, D. K. (1986). *Multitask performance: Predicting success in Naval aviation primary flight training* (NAMRL-1316). Pensacola Air Station, FL: Naval Aerospace Medical Research Laboratory.

Griffin, G. R., & Mosko, J. D. (1977). *A review of naval aviation attrition research (1950-1976: A base for the development of future research and evaluation* (NAMRL 1237), Pensacola Air Station, FL: Naval Aerospace Medical Research Laboratory.

Guilford, J. P., & Lacey, J. I. (1947). Printed Classification Tests. *A.A.F. Aviation Psychological Program Progress Research Report*, 5. Washington, DC: U.S. Government Printing Office.

Harss, C., Kastner, M., & Beerman, L. (1991). The impact of personality and task characteristics on stress and strain during helicopter flight. *The International Journal of Aviation Psychology, 1*(4), 301-318.

Helm, W. R., & Reid, J. D. (2003). Race and gender as factors in flight training success. In *Proceedings of the 45th Annual Conference of the International Military Testing Association* (pp. 123-128), Pensacola, FL.

Helmreich, R. L., Foushee, H. C., Benson, R., & Russini, W. (1986). Cockpit resource management: Exploring the attitude-performance linkage. *Aviation, Space and Environmental Medicine, 57*, 1198-1200.

Hogan, R. T. (1991). Personality and personality measurement. In M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of Industrial & Organizational Psychology* (2nd ed.): Volume 2. Palo Alto, CA: Consulting Psychologists Press.

*Horst, R. L., & Kay, G. G. (1991). CogScreen: Personal computer-based tests of cognitive function for occupational medical certification. In R.S. Jensen (Ed.), *Proceedings of the Sixth International Symposium on Aviation Psychology*. Columbus, OH: The Ohio State University.

Hough, L. M. (1992). The "big five" personality variables – construction confusion: Description versus prediction. *Human Performance, 5*, 139-155.

Hough, L. M. (1998). Personality at work: Issues and evidence. In M. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection*. Mahwah, NJ: Lawrence Erlbaum & Associates.

Hough, L. M. (2001). I/Owes its advances to personality. In B. W. Roberts, & R. Hogan (Eds.), *Personality Psychology in the workplace*. Washington, DC: American Psychological Association.

Hough, L. M., Barge, B., & Kamp, J. (2001). Assessment of personality, temperament, vocational interests, and work outcome preferences. In J. P. Campbell, & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum & Associates.

Hough, L. M., Eaton, N. L., Dunnette, M. D., Kamp, J. D., & McCloy, R. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities [Monograph]. *Journal of Applied Psychology, 75*, 581-595.

Hough, L. M., & Ones, D. S. (2002). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In H. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *International Handbook of Industrial, Work and Organizational Psychology*. Newbury Park, CA: Sage Publications.

*Howse, W. R. (1995, November). *Personality factors in Army aircrew selection* (unpublished handout). Fort Rucker, AL: US Army Research Institute Rotary Wing Aviation Research Unit.

Hunter D. R. (1989). Pilot selection. In M. F. Wiskoff, & G. M. Rampton (Eds.), *Military personnel measurement: Testing, assignment, evaluation.* New York: Praeger.

Hunter, D. R., & Burke, E. F. (1992). *Meta analysis of aircraft pilot selection measures* (ARI Research Note 92-51). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hunter, D. R., & Burke, E. F. (1994). Predicting aircraft pilot-training success: A meta-analysis of published research. *The International Journal of Aviation Psychology, 4*, 297-313.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96* (1), 72-98.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting for error and bias in research findings.* Newbury Park, CA: Sage Publications.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869-879.

*Intano, G. P., & Howse, W. R. (1991). *Predicting performance in Army aviation primary flight training* (ARI Research Note 92-06). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

*Intano, G. P., & Howse, W. R. (1992). Predicting performance in Army aviation flight training. *Proceedings of the 36th Annual Meeting of the Human Factors Society* (pp. 907-911).

*Intano, G. P., Howse, W. R., & Lofaro, R. J. (1991a). *Initial validation of the Army aviator classification process* (ARI Research Note 91-38). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

*Intano, G.P., Howse, W.R., & Lofaro, R.J. (1991b). *The selection of an experimental test battery for aviator cognitive, psychomotor abilities and personal traits* (ARI Research Note 91-21). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*, 371-388.

Jensen, A. R. (1993). Test validity: g versus "Tacit Knowledge." *Current Directions in Psychological Science, 2*, 9-10.

*Jessup, G., & Jessup, H. (1971). Validity of the Eysenck Personality Inventory in pilot selection. *Occupational Psychology, 45*, 111-123.

Keil, C. T., & Cortina, J. M. (2001). Degradation of validity over time: A test and extension of Ackerman's model. *Psychological Bulletin, 127*, 673-697.

*Kilcullen, R., Goodwin, J., Chen, G., Wisecarver, M., & Sanders, M. (undated). *Identifying agile and versatile officers to serve in the objective force.* Unpublished manuscript.

*Kilcullen, R. N., Mael, F. A., Goodwin, G. F., & Zazanis, M. M. (1999). *Predicting US Army Special Forces Field Performance.* Unpublished manuscript.

*Kilcullen, R. N., White, L., Sanders, M., & Hazlett, G. (2003). *Assessment of Right Conduct (ARC) Administrator's Manual.* Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

*King, R. E., & Flynn, C. F. (1995). Defining and measuring the "right stuff": Neuropsychiatrically enhanced flight screening. *Aviation, Space, and Environmental Medicine, 66*(10), 951-956.

*King R. E., Retzlaff, P. D., & McGlohn, S. E. (1997). Female United States Air Force Pilot Personality: The new right stuff. *Military Medicine, 162*, 695-697.

*King, R. E., Retzlaff, P. D., & Orme, D. R. (2001). *A comparison of US Air Force pilot psychological baseline information to safety outcomes* (AFSC-TR-2001-0001). Kirtland AFB, NM: Air Force Safety Center.

*Koonce, J. (1998). Effects of individual differences in propensity for feedback in the training of ab initio pilots. *Proceedings of the 42nd Annual Meeting of the Human Factors Society.*

*Koonce, J., Moore, S., & Benton, C. J. (1995). Initial validation of a basic flight instruction tutoring system (BFITS), In R. S. Jensen (Ed.), *Proceedings of the Eighth International Symposium of Aviation Psychology* (Vol. 2, pp. 1037-1040). Columbus, OH: The Ohio State University.

*Lambirth, T. T., Dolgin, D. L., Rentmeister-Bryant, H. K., & Moore, J. L. (2003). Selected personality characteristics of student naval pilots and student naval flight officers. *International Journal of Aviation Psychology, 13*(4), 415-427.

Lee, W. C., & Drasgow, F. (2003). Using decision tree methodology to predict attrition with the AIM. In *Proceedings of the 45th Annual Conference of the International Military Testing Association* (pp. 310-316), Pensacola, FL.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479-1498.

Lubinski, D., & Dawis, R. V. (1992). Attitudes, skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (2nd Ed., Vol. 3, pp. 1-59). Palo Alto, CA: Consulting Psychologists Press.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Maitland, S. B., Intrieri, R. C., Schaie, K. W., & Willis, S. L. (2000). Gender differences and changes in cognitive abilities across the adult life span. *Aging, Neuropsychology, and Cognition, 7* (1), 32-53.

Martinussen, M. (1996). Psychological measures as predictors of pilot performance. A meta-analysis. *The International Journal of Aviation Psychology, 6*, 1-20.

*Martinussen, M., & Torjussen, T. (1998). Pilot selection in the Norwegian Air Force: A validation and meta-analysis of the test battery. *The International Journal of Aviation Psychology, 8*, 33-45.

Mashburn, N. C. (1934). Mashburn automatic serial action apparatus for detecting flying aptitude. *Journal of Aviation Medicine, 5*, 155-160.

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology, 79,* 493-505.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335-354.

McHenry, J. J., & Rose, S. R. (1988). *Literature review: Validity and potential usefulness of psychomotor ability tests for personnel selection and classification* (ARI Research Note 88-13). Alexandria, VA: US Army Research Insitute for the Behavioral and Social Sciences.

Melton, A. W. (1947). Apparatus tests. *A.A.F. aviation psychology research report, 4.* Washington, DC: U.S. Government Printing Office.

*Morrison, T. R. (1988). *Complex visual information processing: A test for predicting Navy primary flight training success* (NAMRL-1338). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

*Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance, 11*, 145-165.

Mount, M. K., Witt, L. A., & Barrick, M. R. (2000). Incremental validity of empirically keyed biodata scales over GMA and the Five Factor personality constructs. *Personnel Psychology, 53*, 299-323.

*Musson, D. M., Sandal, G. M., & Helmreich, R. L.. (2004). Personality characteristics and trait clusters in final stage astronaut selection. *Aviation, Space and Environmental Medicine, 75*(4), 342-349.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal & Social Psychology, 66*, 574-583.

North, R. A., & Griffin, G. R. (1977). *Pilot selection 1919-1977* (NAMRL Special Report 77-2). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than g. *Journal of Applied Psychology, 79* (6), 845-851.

Ones, D. S., & Viswesvaran, C. (1998). Gender, age, and race differences on overt integrity tests: Results across four large-scale applicant data sets. *Journal of Applied Psychology, 83*, 35-42.

Ones, D. S., & Viswesvaran, C. (2001a). Integrity tests and other criterion-focused occupational scales (COPS) used in personnel selection. *International Journal of Selection and Assessment, 9*, 31-39.

Ones, D. S., & Viswesvaran, C. (2001b). Personality at work: Criterion-focused occupational personality scales used in personnel selection. In B.W. Roberts, & R. Hogan (Eds.), *Personality in the workplace*. Washington, DC: American Psychological Association.

Ones, D.S., Viswesvaran, C., & Schmidt, F.L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance [Monograph]. *Journal of Applied Psychology, 78*, 670-703.

*Operational Psychology Department. (20 July, 2004). *2nd ASTB Workshop* [Briefing Slides]. Naval Air Station Pensacola, FL: Naval Operational Medicine Institute, Naval Aerospace Medical Institute.

*Oppler, S. H., McCloy, R. A., Peterson, N. G., Russell, T. L., & Campbell, J. P. (2001). The prediction of multiple components of entry-level performance. In J. P. Campbell, & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification.* Mahwah, NJ: Lawrence Erlbaum Associates.

Organ, D. W. (1994). Organizational citizenship behavior and the good soldier. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology, 48*, 775-802.

*Pelchat, D. (1997) *The Canadian Automated Pilot Selection System (CAPSS): Validation and cross-validation results.* Paper presented at the Ninth International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

*Pettitt, M. A., & Dunlap, J. H. (1995). *Psychological factors that predict successful performance in a professional pilot program.* Paper presented at the Eighth International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Phillips, H. L., Arnold, R. D., & Fatolitis, P. (2003). Validation of an unmanned aerial vehicle operator selection system. In *Proceedings of the 45th Annual Conference of the International Military Testing Association* (pp. 129-139), Pensacola, FL.

*Portman-Tiller, C. A., Biggerstaff, S., & Blower, D. (1998). Relationship between the aviation selection test and a psychomotor battery. *Proceedings of the 40th Annual Conference of the International Military Testing Association*, Pensacola, FL, USA.

Ree, M. J. (2003). *Test of Basic Aviation Skills (TBAS): Scoring the tests and compliance of the tests with standards of the American Psychological Association* (unpublished technical report). San Antonio, TX: Operational Technologies Corporation.

Ree, M. J. (2004a). *Making scores equivalent for TBAS and BAT* (unpublished technical report). San Antonio, TX: Operational Technologies Corporation.

Ree, M. J. (2004b). *Reliability of the Test of Basic Aviation Skills (TBAS)* (unpublished technical report). San Antonio, TX: Operational Technologies Corporation.

Ree, M. J. (2004c). *Test of Basic Aviation Skills (TBAS): Incremental validity beyond AFOQT Aviator composite for predicting pilot criteria* (unpublished technical report). San Antonio, TX: Operational Technologies Corporation.

Ree, M. J., & Carretta, T. R. (1992). *The correlation of cognitive and psychomotor tests* (AL-TP-1992-0037). Brooks AFB, TX: Armstrong Laboratory, Air Force Materiel Command.

Ree, M. J., & Carretta, T. R. (1996). Central role of g in military pilot selection. *The International Journal of Aviation Psychology, 6*, 111-123.

Ree, M. J., & Carretta, T. R. (1998). Computerized testing in the United States Air Force. *International Journal of Selection and Assessment, 6*(2), 82-89.

*Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior job knowledge in complex training performance. *Journal of Applied Psychology, 80* (6), 721-730.

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1*, 86-89.

Ree, M. J., & Earles, J. A. (1993). g is to Psychology what carbon is to Chemistry: A reply to Sternberg and Wagner, McClelland, and Calfee. *Current Directions in Psychological Science, 2*, 8-9.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-62.

Retzlaff, P. D., King, R. E., & Callister, J. D. (1995). *USAF pilot training completion and retention: A ten year follow-up on psychological testing* (AL/AO-TR-1995-0124). Brooks AFB, TX: Armstrong Laboratory, Air Force Materiel Command.

*Retzlaff, P. D., King, R. E., Callister, J. D., Orme, D. R., & Marsh, R. W. (2002). The Armstrong Laboratory Aviation Personality Survey: Development, norming, and validation. *Military Medicine, 167*(12), 1026-1032.

*Retzlaff, P. D., King, R. E., McGlohn, S. E., & Callister, J. D. (1996). *The development of the Armstrong Laboratory Aviation Personality Survey* (ALAPS) (AL/AO-TR-1996-0108). Brooks AFB, TX: Armstrong Laboratory, Air Force Materiel Command.

Robertson, I. T., & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational Psychology, 66*, 225-244.

Rogers, D. L., Roach, B. W., & Short, L. O. (1986). *Mental ability testing in the selection of Air Force officers: A brief historical overview* (AFHRL-TP-86-23). Brooks Air Force Base, TX: U.S. Air Force Human Resources Laboratory.

*Roscoe, S. N., Corl, L., & LaRoche, J. (2001). *Predicting human performance*. Saint-Laurent, Quebec: Helio Press.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S, III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54* (2), 297-330.

Russell, C. J., Mattson, J., Devlin, S. E., & Atwater, D. (1990). Predictive validity of biodata items generated from retrospective life experience essays. *Journal of Applied Psychology, 75*, 569-580.

Russell, T. L., & Peterson, N. G. (2001). The Experimental Battery: Basic attribute scores for predicting performance in a population of jobs. In J. P. Campbell, & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum Associates.

Russell, T. L., Reynolds, D. H., & Campbell, J. P. (Eds.) (1994). *Building a joint-service classification research roadmap: Individual differences measurement* (AL/HR-TP-1994-0009). Brooks AFB, TX: Armstrong Laboratory.

Saad, S., & Sackett, P. R. (2002). Investigating differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology, 87* (4), 667-674.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. *American Psychologist, 56* (4), 302-318.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49* (11), 929-954.

Salgado, J. F. (1998). Big Five personality dimensions and job performance in Army and civil occupations: A European perspective. *Human Performance, 11*, 271-288.

Schmidt, F. L., & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. *Current Directions in Psychological Science, 2*, 11-12.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82*, 719-730.

Schmitt, N., Sackett, P. R., & Ellingson, J. E. (2002). No easy solution to subgroup differences. *American Psychologist, 58* (4), 305-306.

*Shipley, B. D. (1983). *Maintenance of Level Flight in a UH-1 Flight Simulator as a Predictor of Success in Army Flight Training.* Unpublished manuscript: Army Research Institute for the Behavioral and Social Sciences.

*Shore, W., & Gould, R. B. (2003). *Developing pilot and navigator/technical composites for the Air Force Officer Qualifying Test (AFOQT) Form S* (unpublished technical report). San Antonio, TX: Operational Technologies Corporation.

*Shull, R. N., & Dolgin, D. L. (1989). Personality and flight training performance. *Proceedings of the 33rd Annual Meeting of the Human Factors Society, vol. 2*, 891-895.

*Shull, R. N., Dolgin, D. L., & Gibb, G. D. (1988). *The relationship between flight training performance, a risk assessment task, and the Jenkins Activity Survey* (Interim NAMRL-1339). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

*Shull, R. N., & Griffin, G. R. (1990). *Performance of several different naval pilot communities on a cognitive/psychomotor test battery: Pipeline comparison and prediction* (Interim NAMRL-1361). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

*Siem, F. M. (1991). *Predictive validity of response latencies from computer-administered personality tests*. Paper presented at the 33rd annual conference of the Military Testing Association, 28-31 October, 1991.

*Siem, F. M. (1992). Predictive validity of an automated personality inventory for Air Force pilot selection. *The International Journal of Aviation Psychology, 2*, 261-270.

*Siem, F. M., Carretta, T. R., & Mercatante, T. A. (1988). *Personality, attitudes and pilot training performance: Preliminary analysis* (AFHRL-TP-87-62). Brooks AFB, TX: Air Force Human Resources Laboratory.

*Skinner, J., & Alley, W. E. (2002). *Air Force Officer Qualifying Test (AFOQT): Form R and S development and norms* (unpublished technical report). San Antonio, TX: Operational Technologies Incorporated.

Skinner, J., & Ree, M. J. (1987). *Air Force Officer Qualifying Test (AFOQT): Item and factor analysis of form O* (AFHRL-TR-86-68). Brooks Air Force Base, TX: U.S. Air Force Human Resources Laboratory.

Sperl, T. C., & Ree, M. J. (1990). *Air Force Officer Qualifying Test (AFOQT): Development of quick score composites for forms P1 and P2* (AFHRL-TR-90-3). Brooks Air Force Base, TX: U.S. Air Force Human Resources Laboratory.

Stark, S. E., Chernyshenko, O. S., & Drasgow, F. (2003). A new approach to constructing and scoring fake-resistant personality measures. In *Proceedings of the 45th Annual Conference of the Military Testing Association* (pp 323-329), Pensacola, FL.

Sternberg, R. J., & Wagner, R. K. (1993). Thinking Styles Inventory. In R. J. Sternberg, *Thinking Styles.* New York: Cambridge University Press.

*Street, D. R., Jr., Dolgin, D. L., & Helton, K. T. (1993). Personality tests in an enhanced selection model. In R. S. Jensen & D. Neumeister (Eds). *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 428-433). Columbus, OH: The Ohio State University.

Street, D. R., Jr., Helton, K. T., & Dolgin, D. L. (1992). *The unique contribution of selected personality tests to the prediction of success in naval pilot training* (NAMRL-1374). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

*Taylor, J. L., O'Hara, R., Mumenthaler, M. S., & Yesavage, J. (2000). Relationship of CogScreen to flight simulator performance and pilot age. *Aviation, Space and Environmental Medicine, 71*(4), 373-380.

*Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

Tirre, W. C. (1997). Steps toward an improved pilot selection battery. In R. F. Dillon (Ed.), *Handbook on testing* (pp.220-255). Westport, CT: Greenwood Press.

Toquam, J. L., Corpe, V. A., & Dunnette, M. D. (1989). *Literature review: Cognitive abilities – theory, history, and validity* (ARI Research Note 91-28). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Tupes, E. C., & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings* (ASTD-TR-61-97). Lackland AFB, TX: Aeronautical Systems Division, Personnel Laboratory.

Turnbull, G. J. (1992). A review of military pilot selection. *Aviation, Space and Environmental Medicine, 63*, 825-830.

Vernon, P. E. (1969). *Intelligence and cultural environment.* London, England: Methuen.

Weiss, E. M., Kemmler, G., Deisenhammer, E. A., Fleischhacker, W. W., & Delazer, M. (2003). Sex differences in cognitive function. *Personality and Individual Differences, 35* (4), 863-875.

*Weissmuller, J. J., Schwartz, K. L., Kenney, S. D., Shore, C. W., & Gould, R. B. (2004). *Recent developments in USAF officer testing and selection.* Paper presented at the 46th Annual Conference of the International Military Testing Association, Brussels, Belgium.

*White, L. A., & Young, M. C. (1998). *Development and validation of the Assessment of Individual Motivation (AIM).* Paper presented at the Annual Meeting of the American Psychological Association, San Francisco, CA.

*White, L. A., Young, M. C., Heggestad, E. D., Stark, S., Drasgow, F., & Piskator, G. (2004). *Development of a non-high school diploma graduate pre-enlistment screening model to enhance the future force.* Paper presented at the 24th Army Science Conference, Orlando, FL. (www.asc2004.com/manuscripts)

*White, L. A., Young, M. C., Heggestad, E. D., Stark, S., Drasgow, F., & Piskator, G. (2005). *Army Tier Two Attrition Screen (TTAS) Update* [Briefing Slides]. Presented to The Manpower Accession Policy Working Group, Monterey, CA.

*White, L. A., Young, M. C., & Rumsey, M. G. (2001). ABLE implementation issues and related research. In J. P. Campbell, & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wise, L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L., & Divgi, D. R. (1992). *Sensitivity and fairness of the Armed Services Vocational Aptitude Battery (ASVAB) Technical Composites* (DMDC Technical Report 92-002). Washington, DC: Defense Manpower Data Center, Personnel Testing Division.

*Woycheshin, D. E. (2002). *CAPSS: The Canadian Automated Pilot Selection System.* Paper presented at the Workshop of the RTO Human Factors and Medicine Panel (HFM), Monterey, CA, USA.

Appendix A

Overview of Aviator Selection Test Batteries

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
Tests Used by US Military						
MultiTrack Test Battery (MTTB) **US Army**			> 1 day	n = app. 3000. Attempted to use MTTB to predict 19 primary grades, setbacks, and attritions. Rs ranged from .29 to .39 for flight grades and .21 to .36 for academic grades. Discriminant analysis showed sign difference between setbacks and others. No differences in multiple discriminant analysis between pass and fails. Prediction of flight grades mainly from complex coordination task. Prediction of academic grades mainly from cognitive DVs.	Currently used by US Army to classify pilot trainees into Attack/Scout or Cargo/Utility aircraft training tracks. Already Computerized. Requires non-standard equipment.	Intano and Howse (1991; 1992); Intano, Howse, & Lofaro (1991a, 1991b)
	Complex Cognitive Assessment Battery (CCAB; devel by ARI)				Probably can be Web-Enabled.	
		Tower				
		Following directions				
		Word Anagrams				
		Logical relations				
		Mark Numbers				

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[MTTB con't]		Information Purchase				
		Number & Words				
		Route planning				
	Porta-Bat; selected tests only					
		complex coordination				
		2 hand coordination				
		mental rotation				
		decision making speed				
		timesharing				
		serial mental arithmetic				
		embedded word				
		Manikin			unknown whether it can be Web-Enabled.	
		Word Knowledge				
	Cockpit Management Attitudes Battery				Can be Web-Enabled.	
		Work and Family Orientation Questionnaire				
		Revised Jenkins Attitude Survey				

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[MTTB con't]		Extended Personal Attributes Questionnaire				
		Cockpit Attitudes Management Questionnaire				
	Complex Coordination/Multi-Tasking Battery (CCMB; NAMRL)				combination of psychomotor and dichotic listening task probably cannot be Web-Enabled.	
		Psychomotor-stick				
		Dichotic listenting				
		Psychomotor and dichotic				
		Psychomotor -stick and rudder				
		Stick, rudder, dichotic listening				
		Stick, rudder, throttle				

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
Aviator Selection Test Battery (ASTB) **US Navy**			2.5 hours	Accdg to Operational Psychology Dept Briefing Slides, PFAR predicted flight grades in and attrition from primary flight training; from same presentation, validity data is summarized for FY98-FY04 -- AQR exhibited validity of .46 for Student Naval Aviators in predicting academic aspects of flight training, and PFAR exhibited validity of .32 for predicting peformance in flight training for USN SNAs, and .21 for USMC SNAs (Ns not provided, likely in the thousands; charts suggests that validity for predicting attrition is in the high teens for SNAs (USN & USMC), using the AQR or the PFAR.	became operational in 1992; accdg to ASTB Workshop Briefing Slides, 1992 version was validated separately for SNAs and SNFOs and was developed to be bias-free for gender/ethnicity with ETS Already Computerized. Can be Web-Enabled. Does not require non-standard equipment.	Operatonal Psychology Dept (July, 2004)
				Compared the paper and pencil version of the ASTB with the computrized version administered via APEX. No significant differences in means, variance, model, or predictive validity. However, N = only 82.		Blower (1998)
		Reading Comprehension			Reading & Math are often combined and called the Math/Verbal subtest.	
		Math	15 min			
		Mechanical Comprehension	10 min			
		Spatial Apperception Test	15 min			

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[ASTB con't]		Aviation and Nautical Knowledge			serves as a maximal measure of interests/motivation to be a pilot	
		Aviation Interests	20 min.		has been dropped	
		biographical inventory			has been dropped	
Air Force Officer Qualification Test (AFOQT) **US Air Force** **US Navy**	Verbal (V), Quantitative (Q), Academic Aptitude(AA), Pilot(P), Navigator/Tech (N)	Form Q, 16 subtests; Form S, 11 subtests. Form S dropped RC, DI, MC, EM, SR	4.5 hours (Q), 3.5 hours (S)	low but significant validity of the Pilot Aptitude Rating as a stand-alone for predicting UPT P/F (R=.168) & UPT Rank (R=.20) (Carretta & Ree, 1994); UPT final grade Form S (R=.34,uncorrected) (Shore & Gould, 2003)	rank includes subjective evaluation. Using regression analysis with test scores rather than the composite produces some increase in predictive validities. Criterion data obtained up to 5 years after testing. Not computerized, except for SDI+ Can be Web-Enabled. Does not require non-standard equipment, except TR test.	Carretta & Ree (1994); Form Q, Glomb & Earles (1997); Form S, Gould & Shore (2002)
		Form S: P=AR, MK, IC, TR + AI. N=VA, AR, MK, BC,TR + GS. V=VA + WK. Q=AR + MK. AA=V + Q. (subtest definitions follow)		predictive validity ranges from .00 to .27 for individual test scores predicting academic test grades; from .00 to .13 for individual test scores predicting check ride scores		Ree, Carretta, and Teachout (1995)
				BAT + AFOQT combined to make PCMS score. N= 676. Correlation between pass/fail and PCMS was 0.34 uncorrected and 0.48 when the criterion variable was corrected for dichotimization	BAT was composed of 5 tests at this time	Carretta (2000)

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[AFOQT con't]				Form 0. n=695 (1982-1984). uncorrected predictive validities from -.01 to .34 for pass/fail. R for pass/fail=.395. Pilot composite with 8 tests had R =.21		Arth, Steuck, Sorrentino, Burke (1990)
	First 8 tests below are used in the Pilot Composite	Verbal Analogies (VA)	9 min			
		Mechanical Comprehension (MC)	23 min		dropped from Form S after factor analyses showed that other subtests adequately measured the cognitive predictor space.	
		Electrical Maze (EM)	13 min		dropped from Form S after factor analyses showed that other subtests adequately measured the cognitive predictor space.	
		Scale Reading (SR)	18 min		dropped from Form S after factor analyses showed that other subtests adequately measured the cognitive predictor space.	
		Instrument Comprehension (IC)	9 min			
		Block Counting (BC)	5 min			

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[AFOQT con't]		Table Reading (TR)	9 min		This subtest has been in use since 1942, and it still accounts for unique variance in prediction of pilot performance. Requires non-standard equipment (Large table).	
		Aviation Information (AI)	9 min			
		Arithmetic Reasoning (AR)	30 min			
		Reading Comprehension (RC)	19 min		dropped from Form S after factor analyses showed that other subtests adequately measured the cognitive predictor space.	
		Data Interpretation (DI)	25 min		dropped from Form S after factor analyses showed that other subtests adequately measured the cognitive predictor space.	
		Word Knowledge (WK)	6 min			
		Math Knowledge (MK)	23 min			
		Rotated Blocks (RB)	15 min			
		General Science (GS)	11 min			
		Hidden Figures (HF)	10 min			

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[AFOQT con't]		Self Description Inventory (SDI) (Form S only)	41 min	Currently considered experimental	This is the "original" Big 5: Conscientiousness, Agreeableness, Neuroticism, Openness, Extroversion + Service & Team Orientation Inventory	Skinner & Alley (2002); Christal, et al, 1997
Basic Aptitude Test (BAT) **US Air Force**					Initial description of BAT in Carretta (1987a). Became operational in 1993. Already Computerized. Requires non-standard equipment, dep on test. Can be Web-Enabled, dep. on test.	
				Only Flying Experience showed decent validity as a stand-alone measure, but adding all of them to the AFOQT enhanced predictive validity for UPT P/F (R=.295) and UPT Rank (R=.333) (Carretta & Ree, 1994)	BAT less predictive than in previous samples. May be caused by prediction to Rank, which contains a subjective evaluation.	Carretta & Ree (1994)
				Composite of AFOQT, BAT, and self reported flight hours correlated 0.23 with class rank and -.11 and -.18 with extra flying hours for T-37 and T-38 training. In this study, the BAT included a risk taking measure that was subsequently deleted.	sample from 1986-1992.	Duke and Ree (1996)
[BAT con't]	Psychomotor				By 1993, psychomotor DVs were combined into a composite	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
		Two-hand Coordination	10 min	test-retest reliabilities range from .650 to .823 for retests from 2 weeks to 6 months on each of the 2 DVs	Probably cannot be Web-Enabled. Requires non-standard equipment.	Carretta, Zelenski, and Ree (1997)
		Complex Coordination	10 min	test-retest reliabilities range from .668 to .827 for retests from 2 weeks to 6 months on each of the 3 DVs	Probably cannot be Web-Enabled. Requires non-standard equipment.	Carretta, Zelenski, and Ree (1997)
		Time sharing--tracking with a choice RT task	30 min	n=212. Timesharing DVs not related to pass/fail or check flight scores. Were related to track selection (classification). test-retest reliabilities range from .474 to .808 for retests from 2 weeks to 6 months on each of the 2 DVs	unusual adaptive variable; combined with choice RT task Probably cannot be Web-Enabled. Requires non-standard equipment.	Carretta (1987d); Carretta, Zelenski, and Ree (1997)
		Item Recognition (Sternberg Memory Search)	20 min	test-retest reliabilities range from .518 to .792 for retests from 2 weeks to 6 months on each of the 2 DVs	Probably cannot be Web-Enabled. Requires non-standard equipment.	Carretta, Zelenski, and Ree (1997)
		Mental Rotation	25 min	Addition of mental rotation scores to AFOQT-Pilot composite only improved prediction of track selection (classification).	angle of rotation not linearly related to RT. Dropped. Requires non-standard equipment. Probably cannot be Web-Enabled.	Carretta (1987c)
		Attitudes Toward Risk	10 min		Dropped. Probably can be Web-Enabled. Does not require non-standard equipment.	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[BAT con't]		Flying Experience (self-reported)	??		Probably can be Web-Enabled. Does not require non-standard equipment.	
		Dot estimation	6 min		Dropped; measure of firmness/quickness of decision making Probably cannot be Web-Enabled. Requires non-standard equipment.	
		digit memory	5 min		Dropped. Probably cannot be Web-Enabled. Requires non-standard equipment.	
		encoding speed	15 min	n=2219. Did improve prediction of AFOQT composite. However, test results did not show normal pattern of data.	Test eventually dropped because of negative weight in regression equation -- may have been a problem with the programming. Probably cannot be Web-Enabled. Requires non-standard equipment.	Carretta, 1988
		immediate/ delayed memory	25 min	n=2219. Did not improve prediction of AFOQT composite	Data did not show normal pattern of results. Dropped -- may have been a problem with the programming. Probably cannot be Web-Enabled. Requires non-standard equipment.	Carretta, 1988

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[BAT con't]		choice reaction time	20 min		Dropped -- may have been a problem with programming. Probably cannot be Web-Enabled. Requires non-standard equipment.	
		embedded figures	15 min	no significant relation to training performance. Did not increase predictive validity of AFOQT pilot composite	Dropped. Probably can be Web-Enabled. Requires non-standard equipment.	Carretta (1987b)
		self-crediting word knowledge	10 min		Eventually dropped because of negative weight in regression equation; intended as a measure of Self-Confidence Probably can be Web-Enabled. Does not require non-standard equipment.	
		activities interest	10 min	test-retest reliabilities range from .655 to .871 for retests from 2 weeks to 6 months on each of the 2 DVs	The Activities Interest scale is a forced-choice instrument that measures risk-taking attitudes in situations that involve threats to physical survival. Can be Web-Enabled. Does not require non-standard equipment.	Carretta, Zelenski, and Ree (1997); Siem, Carretta, & Mercatante (1988)

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
Enhanced Flight Screening Program **US Air Force**	Armstrong Laboratory Aviation Personality Survey (ALAPS)		20-30 min	describes test development. Conversion tables for men and women. Cronbach alphas 0.73-.90. Administered with Neo-PI-R. Except for Openness, high correlation between some of ALPS scales and NEO summary scales. No NEO equivalent for 7 ALAPS scales	Validity data not currently available. Test questions and scoring key published in an Armstrong Laboratory tech report. US Navy intends to evaluate at least some of the scales from this inventory for potential use in their pilot selection process. Has different scoring conversions (essentially norms) for men vs women. Can be Web-Enabled. May require non-standard equipment.	King & Flynn, 1995; Retzlaff, King, McGlohn, & Callister (1996); Retzlaff, King, Callister, Orme and Marsh (2002); Operatonal Psychology Dept (2004)
		5 "personality" scales: confidence, socialness, aggressiveness, orderliness, & negativity				
		6 "crew interaction" scales: Dogmatism, Deference, Team Orientation, Organization, Impulsivity, & Risk-Taking				
		4 "psychopathology" scales: Affective Lability, Anxiety, Depression, & Alcohol Abuse				

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[EFSP con't]	NEO-PI-R	Big 5 scales: Extraversion; Conscientiousness, Agreeableness, Emotional Stability, Openness	~40 minutes	Archival analysis of accident data. Pilots with higher scores on Competence are 4 times more likely to have an accident; those with high scores on Dutifulness are 2 times more likely. (Competence & Dutifulness are two facets of the Conscientiousness domain score.)	General literature shows that the NEO-PI-R is valid for a veriety of jobs, but almost none of the criterion-related validity evidence (to date) is specific to the pilot job; inventory was carefully developed and is psychometrically sound in terms of reliability and construct validity. Does not require non-standard equipment.	King, Retzlaff, and Orme (2001)
				Cluster analysis revealed no domain combinations related to UPT pass/fail. Attempted to analyze only Self-Initiated Terminations (SIEs) but needed larger sample to do odds ratios.	First attempt to form profiles from NEO and calculate likelihoods. Seems promising, especially for SIEs. Must keep in mind, however, that the percentage of SIEs is typically very small, so the low base rate will make it difficult to show that a predictor does better than chance alone.	Anesgart and Callister (1999)
	Multi Attribute Battery (MAB) II				Does not require non-standard equipment.	
		Information				
		Comprehension				
		Arthimetic				
		Similarities				
		Vocabularly				

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[EFSP con't]		Digit Symbol				
		Picture Completion				
		Spatial				
		Picture Arrangement				
		Object Assembly				
	CogScreen (battery)				CogScreen originally developed for the FAA's Civil Aeromedical Institute (CAMI); USAF did some research with it related to pilots, but not for selection. It was eventually dropped by the USAF due to high royalty fees.	
	MicroCog				replaced Cogscreen	
		attention/mental control				
		memory				
		reasoning/calculation				
		spatial processing				
		reaction time				
		information processing accuracy				
		information processing speed				
		cognitive functioning				
		cognitive proficiency				

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
Test of Basic Aviation Skills (TBAS) **USAF & US Navy are studying**				N=531 USAF pilot trainees (already selected on other measures); horizontal tracking (rudder pedals), "airplane" tracking (joystick), and Emergency Scenarios showed low, but significant correlations with training pass/fail and training performance (composite of grades, check rides, etc.). (Multiple-task test did not.) Multiple R for predicting T-37 Total score (check ride, instructor ratings, quiz scores, etc) was .33; Multiple R for predicting UPT pass/fail was .31.	Doesn't look like regressions were corrected for shrinkage but author notes that they were downwardly biased due to a high degree of range restriction. N=126 USAF pilot trainees – test-retest reliability was very low for Emergency Scenario score (.15) and acceptable but not impressive for the other scores (.56-.75) Documentation describing how score variables were derived does not appear to exist, nor does computer programming documentation. Probably cannot be Web-Enabled. Requires non-standard equipment.	Operational Psychology Dept (20 July 2004); Ree (2003, 2004a, 2004b, 2004c)
		Direction orientation			Spatial orientation test.	
		Multi-Tasking (Tracking & Directed Listening tasks performed simultaneously)			A tracking task requiring use of rudder pedals and a joystick, is performed at the same time as a Directed Listening task (press trigger on joystick when any of several pre-specified digits are heard out of a stream of randomly-presented digits).	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[TBAS con't]		Emergency Scenarios			Test-taker must notice and respond quickly to warning signals indicating that an "emergency" is occurring.	
Armed Services Vocational Aptitude Battery (ASVAB) **US Army, US Air Force, US Navy**				In a large sample of enlisted US Army Soldiers, a composite of ASVAB factor scores demonstrated an average multiple correlation, across several MOS (N ranged from 73 to 553 per MOS), of .62 for predicting Core Technical Job Proficiency, .66 for predicting General Soldiering Proficiency, .37 for predicting Effort and Leadership, .17 for predicting Maintaining Personal Discipline, and .16 for predicting Physical Fitness and Military Bearing. The multiple correlations were corrected for range restriction and adjusted for shrinkage.	Currently used as a prescreening tool by US Army for Warrant Officer selection into rotary wing pilot training, and as a prescreening for enlisted personnel in all branches of the US military. Does not require non-standard equipment.	Oppler, McCloy, Peterson, Russell, and Campbell (2001)
		General Science (GS)	11 min			
		Arithmetic Reasoning (AR)	36 min			
		Word Knowledge (WK)	11 min			
		Paragraph Comprehension (PC)	13 min			
		Auto & Shop Information (AS)	11 min			
		Mathematics Knowledge (MK)	24 min			
		Mechanical Comprehension (MC)	19 min			

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[ASVAB con't]		Electronics Information (EI)	9 min			
		Assembling Objects (AO)	9 min		added in 2002 to replace the Coding Speed test; good variance, smaller gender differences	
		Numerical Operations (NO)	--		dropped as of 2002 due to unacceptably large gender differences	
		Coding Speed (CS)	--		dropped as of 2002 due to unacceptably large gender differences	
Tests/ Batteries Used by Foreign Military Services						
MICROPAT **UK ARMY**			Unknown	Good validity as a stand alone predictor of success in Basic & Advanced rotary wing pilot training (shrunken R ~.50); added quite a bit of incremental validity when used in conjunction with a spatial/psychomotor composite & bare-bones biographical info. (Bartram, 1987), n=105.	Bartram sample was rotary wing pilots. Tracking tasks predict basic training outcome, not advanced. Requires non-standard equipment.	Bartram (1987)
	Psychomotor					
		Adaptive Pursuit Tracking			Probably cannot be Web-Enabled.	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[MICROPAT con't.]		Compensatory Tracking			Probably cannot be Web-Enabled.	
	Information Management Ability					
		Risk - figure out risks, which change over time, and adapt strategies accordingly			Almost identical to NAMRL Automated Risk Assessment Task. Both based on Slovic's (1969) paper. May be Web-Enabled, dep on degree of timing accuracy needed.	
		Signal - signal detection			Signal detection theory task. Measure beta, decision speed, and accuracy (d'?). May be Web-Enabled, dep on degree of timing accuracy needed.	
		Dual Tasks (tracking & mental arithmetic)			May be Web-Enabled, dep on degree of timing accuracy needed.	
		Landing (simulated aircraft landing)			superfically resembles an aircraft landing simulator. May be Web-Enabled, dep on degree of timing accuracy needed.	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[MICROPAT con't.]		Scheduling (dynamic speed/distance estimation task)			Resembles POPCORN task used by NASA. Correlates with tests of mathematical reasoning and knowledge. May be Web-Enabled, dep on degree of timing accuracy needed.	
		digit span		n=105. Pass/fail for basic r=.57 and .52 for pass/fail at advanced (shrinkage to .50 and .44). With 16 PF, R=.75 for pass/fail in advanced	digit span included with other tests in this article. Cattell 16 PF also included and by itself r= .57 with pass fail for advanced training. MicroPat with 16 PF has same predictive validity as then current selection test plus bio. Bartram (1987) does not include digit span in MicroPat	Bartram and Dale (1985)
RAF (computer-based testing) **RAF; Australian RAF; Netherlands RAF, British Army Air Corps**			Unknown	Meta analysis. Criterion is pass/fail on first attempt of initial flying training; data from Turkish AF, RAF, and AAC (British rotary wing). ns of about 2000 to 4000. Corrected validity for the average stanine score was .45 for the AAC, .42 for the RAF, and .35 for the Turkish AF	May be Web-Enabled, dep on test. Requires non-standard equipment.	Burke, Hobson, and Linksy (1997)
		Sensory Motor Apparatus			Traditional RAF test. Probably cannot be Web-Enabled. Requires non-standard equipment.	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[RAF con't.]		Instrument Comprehension			Traditional RAF test. Probably can be Web-Enabled. Does not require non-standard equipment.	
		Control of Velocity			Traditional RAF test. Probably cannot be Web-Enabled. Requires non-standard equipment.	
Canadian Automated Pilot Selection System (CAPSS) **Canadian Military**			5 hours	Validities in the .25-.35 range for predicting variables such as Training P/F, Course Grade, Performance Ratings, and Academic Average.	CAPSS scores are pretty highly correlated with previous pilot experience (~.35) Probably cannot be Web-Enabled. Requires non-standard equipment (flight simulator).	Woychesin (2002)
				r=.47 between CAPSS score and pass/fail	non representative cadet sample; no females or French speakers ; 70% pass rate. Cadets prescreened	Pelchat (1997)
Dichotic Listening Test **Israeli Military**			25-35 min, dep on version		Maybe can Web-Enabled. Requires non-standard equipment.	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[DLT con't.]				n=100 cadets and n=95 pilots. Very low error rate (0.7 to 4.3%) in Part 1. Non significant correlation with intelligence test, mechanical and psychomotor test, and psychiatric evaluation. Correlates .26+ with 3-point pass/fail early/late criterion. Second study with n=95 shows differences between transport pilots and jet pilots.	Errors may be caused by speaking response as next stimulus is presented. Test shows rapid learning. 58% failure rate in flight training! Israelis use a cascade selection system, not a multi-hurdle	Gopher and Kahneman (1971)
				Pilot trainees who passed training made fewer errors than those who failed training; test scores added a bit of incremental validity when added to a battery of 15 undescribed subtests (probably cognitive), a psychomotor test, and an interview.		Gopher (1982)
				Dichotic listenting test performed alone and with one, two, or three tracking tasks. N= about 675 student Naval aviators. Dichotic alone had max. possible right of 108; mean was 101.9. After transformation, dichotic scores did not predict pass/fail, but did predict flight grade and were independent of other measures.		Delany (1990)

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
Pilot Evaluation System (PES) **Israeli AF**	22 scenarios with any combination of 5 tasks; figure of merit DV		90 min	Internal validity study at USAF Academy. Then correlated PES scores with daily scores (not described further) and maneuver grades. Seemed promising.	Simulates air-to-air engagements. US company may have bought rights. No other known English-language research. Probably cannot be Web-Enabled. Requires non-standard equipment (static part-task simulator).	Garvin, Acosta, and Murphy (1995)
Pilot Selection Tests not specifically developed by/for the Military						
WOMBAT-CS **Canadian National Flight College; commercial airlines, flight schools**			3.5 hours, but can be shortened if desired		Wombat has been used as a last-stage selection test by the Canadian National Flying College in Quebec for 7 years. Wombat scores are 70% of total score for last-stage selection. Little published validity evidence because of proprietary restrictions. Publisher is working on a web version. Requires non-standard equipment.	Roscoe, Corl, & LaRoche (2001); Aero Innovation (1998)

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[WOMBAT-CS con't.]				Study 1: 32 ab initios. Criteria were academic performance, time to solo, and time to check ride recommendation. Bonus score significantly related to academic performance (r=.36). Study 2: 21 students from Phase 1. Simulated cross-country flight. SA estimates by questionnaire. Bonus score and total score correlated (r<.4) with SA measure.		Frey, Thomas, Walton, and Wheeler (2001)
				Study conducted at aviation dept. of US college. Subjects from all 4 years of program. Wombat scores related to flight hours (r=.35). Author believes WOMBAT can be used to measure SA and stress tolerance.		Cain (2002) dissertation abstract
	total performance					
		Two-Handed Tracking Task (Primary)				
		Figure Rotation (Bonus)				
		Quadrant Location (Bonus)				
		Digit Canceling (Bonus)				
BFITS **Developed under SBIR contract for AF**	several parameters from flight simulator performance		50 hours		computer hardware and software may need to be updated. Cannot be Web-Enabled. Requires non-standard equipment.	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[BFITS Con't]				Study at ComAir Academy (n=72). r=.33 between time to gain private pilot's license and scores in first phase of BFITS (first 5 hours?).	BFITS is a combination training and selection system conducted on PC-based flight simulator. 31 lessons leading to private pilot's license. Each lesson has tutorial with procedural training. DV is trials to criterion performance.	Davis, Koonce, Herold, Fedor, and Parsons (1997)
				n=10 in experimental group; n=17 in control group. BFITS students soloed after 12.5 hours versus 17.6 for the control (sign at <01).	initial validation study on civilian pilot candidates	Koonce, Moore, and Benton (1995)
				181 total cadets at ComAir Academy. No data provide but authors claim same pattern as 1997 study.		Koonce (1998)
PILAPT			varies; long and short version availabl e		commercially available product. May be as expensive as WOMBAT with additional per usage fees. Tests added and deleted over time. Probably cannot be Web-Enabled. Requires non-standard equipment.	
		Sequences	8 min			
		Patterns	10 min			
		Digits	??			
		Concentration	8 min			
		Rules	??		not in current battery	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[PILAPT Con't]		Views	20 min		not in current battery	
		Plans	??		spatial test; not in current battery	
		Shapes	??		spatial test; not in current battery	
		Gates	??		2-d compensatory tracking; not in current battery	
		Ascent	??		job sample; not in current battery	
		In-Flight	??		job sample; not in current battery	
		Descent	??		job sample; not in current battery	
		Deviation Indicator	7 min	unselected sample of RAF University Air Squadron applicants. R based on 169 cadet = .55. Criterion was average training grade after 6 months. Another paper indicates that a composite PILAPT score increased 0.8 SD on retest in this study	4 month test-retest =0.80 n=109	Burke, Kitching, and Valser (1997)
		Trax	5 min		3-d pursuit tracking; 4 month test-retest =0.84 n=109	
	speed; accuracy; combination of two	Hands	10 min		also loads on a spatial factor; 4 month test-retest =0.69 n=109; Spearman-Brown >.91 all DVs	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[PILAPT Con't]		Capacity	15 min.		Not used in ab initio selection. Primary task with two secondary tasks. Single, dual, and triple task performance assessed	
Tests Tried by Military But Not Currently Used						
Cognitive Laterality Battery (CLB)			~ 90 minutes	Visuospatial subtests differentiated pilot trainees who completed training from those who dropped out; verbosequential did not; don't know if test score added incremental validity beyond ASTB scores	study conducted with US Navy pilot trainees, some of whom were rotary wing pilots. Maybe can be Web-Enabled. Requires non-standard equipment.	Gordon & Leighty (1988)
	visuospatial					
		Localization				
		Orientation (mental rotation)				
		Touching Blocks				
		Form Completion				
	verbosequential					
		Serial Sounds				
		Serial Numbers				
		Word Production, Letters				
		Word Production, Categories				

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
Computer-Based Performance Test (CBPT or CBT) **US Navy is studying**			Unknown	n=372 Shull and Griffin (1990) used an early version of the CBT and examined different groups of aviators (2 jet groups, helos and student naval aviators). Helo pilots never did the best on any DV and were often the worst and performed more poorly than the students. Among the students, the ones who eventually went to helos did more poorly than the other groups.	Street, Dolgin, & Helton (1993) describe 4 CBT tasks, as shown in this table. Portman-Tiller, et al. (1998) describes the CBT has having 10 tasks:stick task; dichotic listening; stick and dichotic listening,stick and rudder task; stick, rudder, and dichotic listening; stick, rudder, and throttle;horizontal tracking; absolute difference; mannikin; and absolute difference and horizontal tracking. Probably cannot be Web-Enabled. Requires non-standard equipment.	Delaney (1990); Portman-Tiller, et al. (1998); Shull & Griffin (1990)
				n=106 Hierarchical regression found that all four CBT tasks significantly increased the R. Three scales from the Pilot Personality Questionnaire (assertiveness, competitiveness, and self-control) also significantly increased the R. Analysis focused on predicting advanced naval flight training (not primary or intermediate).	n = 150. factor analysis of CBT and 4 tests of ASTB (no bio inventory) produces 4 factor solution. Dichotic listening task (and combinations) define one factor; stick and combinations define another factor	Street, Dolgin, and Helton (1993)

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[CBPT Con't]					Both Shull and Griffin (1990) and Delaney (1990) describe the transformation needed for the dichotic listening scores--Log to the base 10(#errors +1). They both mention extreme skewness. Given the limited information provided, it appears that many applicants make no errors and a few make many--neurological test?	
		psychomotor/dichotic listening		-0.36 uncorrected predictive validity with Advanced Flight Grade (AFG)		Street, Dolgin, and Helton (1993)
		absolute difference and horizontal tracking		-0.25 uncorrected predictive validity with Advanced Flight Grade (AFG)		Street, Dolgin, and Helton (1993)
		complex visual task		-0.17 uncorrected predictive validity with Advanced Flight Grade (AFG)	based on task analysis of skills needed to extract information from visual displays (Robertson and Castle, 1996)	Street, Dolgin, and Helton (1993)
		mannikin		-0.29 uncorrected predictive validity with Advanced Flight Grade (AFG)		Street, Dolgin, and Helton (1993)

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
Complex Visual Information Processing Test **US Navy is studying**			60-75 min	n=451 student naval aviators undergoing primary training. Number correct r= -.27 with pass/fail. For students who pass training, correlation with each of the 3 DVs and flight grade was significan but low (r = .10). However, number correct exhibited significant correlation with all measures of primary flight training, including academic grades. Stepwise regressions to predict Pass/Fail, Composite Score, and Flight Grade were predicted by number correct, RTs, and biographical inventory, R= .26 to .32. FAR scores and AQT scores never entered the equation.	DVs are number correct, time to read the question, and time to respond. Probably cannot be Web-Enabled.	Morrison (1988)
				Student naval aviators, n ranges from1077 to 337. Test battery. After controlling for intelligence, CVT accounted for about 3.5% of variance in pass/fail. Interacted with accession source and college major. Best linear model from battery included 7 demographic variables, 6 AQT/FAR variables, 3 CVT scores, and 12 interactions; still only accounted for 8.3% of variance in pass/fail.	psychomotor/dichotic listening task did not add variance to the model.	Blower and Dolgin (1990)

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
Multi Attribute Battery (MAB) **NASA uses for astronaut selection**			70 min (7 min per subtest)		better than WAIS for discriminating at high end of scale. Evidence of ceiling effects on some scales. In astronaut population, may have 3 factor solution, not two. Does not require non-standard equipment. Developer refuses to be Web-Enabled.	Bishop, Faulk, Santy (1996)
					highly regard by USAF	
	Verbal IQ (crytallized intelligence)					
		Information				
		Comprehension				
		Arithmetic				
		Similarities				
		Vocabulary				
	Performance IQ (fluid intelligence)					
		Digit symbol				
		Picture completion				
		Spatial				

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[MAB Con't]		Picture arrangement				
		Object assembly				
	full scale IQ (comprised of all subtests)					
CogScreen (Aeromedical Edition) **US Navy is studying;** **USAF considered**	most tasks have speed, accuracy, and throughput score		45 min?	no predictive validity evidence available that we can find. Being considered by the Navy as a pilot selection instrument.	list of tasks and what they measure differs in King and Flynn (1995) versus Callister, King, and Retzlaff (1996), who also give DVs. Callister et al. note that many of the mean accuracy scores are 90+%. Believe that software has scoring error in calculating scores on Pathfinder. Also some of the 65 variables are of questionable use. Probably cannot be Web-Enabled. Requires non-standard equipment.	Callister, King, and Retzlaff (1996); best description of tasks in Horst and Kay (1991)
					study of factor structure of CogScreen by USAF (Carretta, Ree, and Callister, 1999). CogScreen does not appear to be g loaded. Different study -- n=1015 pilot applicants. Data do not fit the original 9 factor Kay model. No obvious single higer-order factor. Best model has 2 higher order factors and 6 first levels	Carretta, Ree, & Callister (1999)

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
[CogScreen con't]					study of aging, flight perforamnce, and Cog Screen shows a 5 factor solution. However, only used 28 of the 65 variables	Taylor, O'Hara et al. (2000)
	1 DV	Backward Digit Span				
	3 DVs	math problems				
	3 DVs	visual sequence comparison				
	4 DVs	symbol digit coding				
		symbol digit coding-immediate recall			listed in King and Flynn only	
	3 DVs	matching to sample				
	3 DVs	mannikin figures				
	7 DVs	divided attention				
	3 DVs	auditory sequence comparison				
	12 DVs	pathfinder				
		symbol-digit coding--delayed recall			King and Flynn only	
	16 DVs	shifting attention			Callister, King, and Retzlaff only; sounds very similar to the Wisconsin Card Sorting Task, which measures perseveration	
	10 DVs	dual-task test			Callister, King, and Retzlaff only	

Battery Name	Composite Score/Test	Subtest	Length	Summary of Validity Evidence	Comments	Key Reference(s) Describing Test or Documenting Validity
PASS			5 hours	223 officers and 232 warrant officer candidates. All qualified for flight training on basis of FAST scores. Predictive validity apparently about .50.	Job sample test. Five 1-hour sessions in UH-1 instrument flight trainer. Data obtained in 1978 and 1979. Analysis based only on data from first session, straight and level flight. Data anlaysis complex and costly. Simulator costs and analysis costs too high; PASS not cost effective. Cannot be Web-Enabled. Requires non-standard equipment (sim).	Shipley (1983)
Complex Coordination Test **US Army USAF (may use)**				n = 1123. r = .50 for pass/fail flying only; about 0.29 for pass/fail for all reasons	Included in preliminary FAST validation. Probably cannot be Web-Enabled. Requires non-standard equipment.	Shipley (1983)
Rotary Pursuit **US Army**				n=1112. r =.30 for pass/fail flying only; about .25 for pass/fail for all reasons	Included in preliminary FAST validation. Probably cannot be Web-Enabled. Requires non-standard equipment.	Shipley (1983)

Appendix B
Overview of Non-Cognitive Inventories that may be Relevant for Aviator Selection

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Inventories Considered or Used for Pilot Selection						
Self Description Inventory, Plus (SDI+) US Air Force	Big 5: Neuroticism; Extraversion; Openness; Agreeableness; Conscientiousness; PLUS Service Orientation and Team Orientation 220 items; "behavior statements" with Likert-type responses 41 min	D: Rational and factor-analytic S: Rational	in progress	No criterion-related validity evidence for the SDI+; should be available in December 2006. A small pilot test (N= 71 enlisted airmen) based on the original 163-item version of the SDI (w/o Service or Team Orientation) showed raw correlations of .47 (Agreeableness), .20 (Conscientiousness), .01 (Extroversion), -.24 (Neuroticism), and .35 (Openness) with a composite supervisory rating variable called "Global 2: Interpersonal Proficiency." The validity coefficients for Agreeableness, Neuroticism, and Openness are statistically significant at $p<.05$. The 5-factor model was originally developed using a sample of USAF enlisted personnel, but the factor structure was confirmed in a sample of 523 USAF commissioned officers, using a precursor of the SDI+.	Skinner & Alley (2002); Christal, et al, 1997; Weismuller, et al., 2004	163-item version of the SDI was originally developed by Christal and colleagues, using trait labels and "behavioral statements" to create items that reference the Big 5 personality factors. The items are not the same as the items in the NEO-PI. The USAF decided to write and add scales that measure Team Orientation and Service Orientation, and called the resulting inventory the SDI+. When they did this, they converted all items to "behavioral statements." The SDI is often listed as part of Form S of the AFOQT, but is currently considered experimental. Computerized.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Self Description Inventory, Plus (SDI+) (Cont.)						Examined test-retest reliability in the 163-item version in a sample of 584 enlisted USAF personnel across 24 bases. The retest interval ranged from 13-25 months. After removing ~20 outliers on each scale, the stability correlation was .65 for Agreeableness, .70 for Conscientiousness, .69 for Extroversion, .79 for Neuroticism, and .72 for Openness. When the sample was grouped into interval ranges, the test-retest tended to be slightly higher for those who retested after a shorter interval (still at least 13 months) than those who retested after a longer interval.
						The UK Defence Research Agency has also conducted research on the SDI, in a military officer sample, using the original SDI. Found the same 5-factor structure.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Noncognitive Scales included in the Basic Aptitude Test (BAT) US Air Force	at least 5 different non-cognitive tests have been tried out, but most were eventually dropped; at present, it sounds like Flying Experience and, maybe, Activities Interest are still part of the battery. presumably biodata for Flying Experience; "activities" for Activities Interest	Unknown	Yes	N= 678 pilot trainees. Carretta & Ree (1993) found that both Flying Experience and the Activities Interest scale (labeled "Risk-Taking" in this study) exhibited a significant correlation with UPT P/F (R=.167 and .101 for Fly Exp and Act Int, respectively) and UPT Rank (R=.190 and .108 for Fly Exp & Act Int, respectively). A surprise finding was that Flying Experience alone correlated almost as high as the AFOQT score alone with these two criterion measures. When Flying Experience was added to the AFOQT, R increased by .067 for UPT P/F (from .168 to .235) and by .074 for UPT Grade (from .168 to .274). When Activities Interest was added to the AFOQT, R increased by .035 for UPT P/F (from .168 to .203) and by .036 for UPT Grade (from .168 to .236).	Carretta & Ree (1993) [NOTE: Carretta & Ree (1994) may include the same data.	The Activities Interest scale is a forced-choice instrument that measures risk-taking attitudes in situations that involve threats to physical survival. Computerized.
NEO-PI US Air Force; NASA; (also used widely in private sector)	Personality: Neuroticism; Extraversion; Openness; Agreeableness; Conscientiousness 240, 5-option items + 3 response validity items, takes ~40 minutes; NEO-FFI has 60, 5-response items	D: Rational & Factor-Analytic S: Rational	yes, but not operationally	General literature shows that the NEO-PI-R is valid for a variety of jobs, but none of the criterion-related validity evidence (to date) is specific to the pilot job; inventory was carefully developed and is psychometrically sound in terms of reliability and construct validity.	www.rpp.on.ca/neopir.htm	Six facet scales underlie each of the 5 summary scales. Believe that the NEO-PI would be available free-of-charge to the US Army. However, Big 5 measure may not cover all of the non-cognitive traits most important for Army Aviators. Contains several validity items, but no "faking" scale per se. Computerized.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
NEO-PI (Cont.)				Compared 48 female AF transport and tanker pilots to 64 comparable male pilots and 103 female undergraduates on the NEO-FFI. Female pilots had significantly higher scores on the Extraversion, Agreeableness, and Conscientiousness scales than male pilots. Also showed large differences from controls (female undergraduates) on same three scales plus Neuroticism.	King, Retzlaff, and McGlohn (1997)	
				scale norms are available for US Air Force student aviators (N=1031; 103 of them females) and for US Army student aviators (N=355) and active duty aviators (N=255); could potentially be used for profile matching purposes	Callister, King, Retzlaff, & Marsh (1999); Howse (1995)	Callister, et al. provide separate norms for males & females, which could be problematic in a selection setting; in their sample (with only 103 females), females scored ~.60 SDs higher on Neuroticism, ~.50 SDs higher on Openness; ~.33 SDs higher on Agreeableness, and not much diff on Extraversion or Conscientiousness (these findings are at least consistent in direction with male-female diffs in the general population for Neuroticism and Agreeableness); also note that some of the students undoubtedly did NOT succeed in becoming an aviator. Howse didn't indicate the number of females, if any, in his sample.
				differences between male and female aviation students disappeared by the third year. 8 of the 36 facets showed significant correlations with GPAs	Pettitt and Dunlap (1995)	

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
NEO-PI (Cont.)				no significant differences on any scale between candidates accepted or rejected for astronaut positions	Musson, Sandal, Helmreich (2004)	
				N=1031 individuals accepted into US Navy's Enhanced Flight Screening program. Cluster analysis revealed that no combination of domain (I.e., Big 5 dimension) scores were correlated with completion of the training program (88% completed). The authors also outlined a methodology for calculating the odds (based on logistic regression) that a pre-trainee with a particular NEO profile will behave like a benchmark profile (e.g., a stellar performer). Preliminary analyses suggests there might be some utility here, e.g., for predicting Self-Initiated Terminations (SIEs) but sample sizes would need to be much larger to go very far with this approach.	Anesgart and Callister (1999)	First attempt to form profiles from NEO and calculate likelihoods. Promising, especially for self-initiated terminations.
				Archival analysis of accident data. Pilots with higher scores on Competence are 4 times more likely to have an accident; those with high scores on Dutifulness are 2 times more likely. (NOTE: Competence & Dutifulness are facet scores within the Conscientiousness factor.)	King, Retzlaff, and Orme (2001)	These results are very counterintuitive.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Armstrong Laboratory Aviation Personality Scale (ALAPS) US Air Force	5 "personality" scales: Confidence (Narcissism), Socialness, Aggressiveness, Orderliness (Compulsivity), & Negativity (Passive-Aggression) 240 True-False items; 20-30 minutes	D: Rational S: Rational	Experimental	no criterion-related validity evidence until at least Dec 2006	Retzlaff, King, McGlohn, & Callister (1996); Retzlaff, King, Callister, Orme, & Marsh (2002); Operational Psychology Dept (20 July 2004)	Test questions and scoring key published in Armstrong Laboratory tech report. US Navy intends to evaluate at least some of the scales from this inventory for potential use in their pilot selection process. Has different scoring conversions (essentially norms) for males and females. Computerized.
	6 "crew interaction" scales: Dogmatism (Authoritarianism), Deference (Submissiveness), Team Oriented, Organization, Impulsivity, & Risk-Taking					In original development study, ALAPS administered to 5,000+ pilot trainees to establish norms and examine gender differences; Females scored ~1.5 SDs higher on Affective Lability and ~.5 SDs lower on Confidence & Dogmatism. N=1,131 --> scale alphas range from 0.73 to .90, and correlations between ALAPS and NEO-PI summary scales were fairly high and logically consistent, except that no ALAPS scales correlated very highly with NEO Openness. There was no NEO equivalent for 7 of the ALAPS scales.
	4 "psychopathology" scales: Affective Lability, Anxiety, Depression, & Alcohol Abuse					USAF intends to administer it to pilot trainees after acceptance into training, as a screening device.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Personal Characteristics Inventory (PCI) NASA; US Air Force; Commercial aviation	Personality: Instrumentality (pos & neg); Expressivity; Verbal Aggressiveness; Mastery Orientation; Work Orientation; Competitiveness; Achievement Striving; Impatience/Irritability; Negative Communion 225, 5-response items, ~ 40 minutes	D: Rational, in that scales selected to measure specific constructs S: Likely some of each	yes, but not operationally	n=259. No PCI or NEO-FFI scales distinguished between applicants selected versus rejected for the astronaut program (analyses conducted separately for males & females). About 1/3rd of the astronaut applicants was classified into each of 3 PCI clusters; the percentage falling in each cluster did not differ between successful (N=63) and non-successful applicants (i.e., persons who ultimately were selected as an astronaut). Musson, et al. say that similar clusters have been found to be predictive of performance and cite some references, two of which sound like they involve pilots or pilot trainees.	Musson, Sandal, and Helmreich (2004); King & Flynn (1995)	The PCI was administered as part of USAF's Enhanced Flight Screening program (King & Flynn). Accdg to Musson, et al., it consists of 11 scales taken from pre-existing instruments – the Extended Personal Attributes Questionnaire, the Work and Family Orientation Questionnaire, and the Jenkins Activity Survey. These authors examined only 10 of the scales. King & Flynn say the PCI was developed by Helmreich & colleagues at the University of Texas - Austin and has been used as a CRM measurement tool by the USAF and commercial aviation. They say that 20% of the item content overlaps with the NEO-PI, but suspect they were referring to correlations with the NEO scales, not strict item overlap.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Personal Characteristics Inventory (PCI) **(Cont.)**				N=23, three-man commercial airline crews. Examined performance of crews, separating them into comparison groups according to the pilot's score on one of 3 clusters (Positive Instrumentality/Expressivity - "Right Stuff", Negative Instrumentality - "Wrong Stuff", or Negative Expressivity - "No Stuff". NOTE: Cluster configurations may not be exactly the same as in later research. Found that the crews with a "Right Stuff" pilot consistently performed effectively, although not necessarily always better than crews with a "Wrong Stuff" pilot.	Chidester & Foushee (1991)	Musson, Sandal, & Helmreich (2004) say the PCI aims to capture positive & negative aspects of two core dimensions of personality: instrumentality & expressivity. Instrumentality traits refer to overall goal seeking and achievement motivation while expressive traits are related to interpersonal sensitivity and concern. Musson, et al. cluster analyzed the PCI scale scores, starting with clusters that had been identified in previous research. Three clusters emerged as in prior research (Gregorich, et al., 1989): Positive Expressive/Instrumental ("Right Stuff"), Negative Instrumentality ("Wrong Stuff"), and Low Motivation ("No Stuff").

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Personal Characteristics Inventory (PCI) **(Cont.)**				N=538 US military officers in transport operations who held positions on the flight deck, and who attended a training program in cockpit resource management. Authors used cluster analysis to identify 3 clusters on the PCI which they labeled "Right Stuff," "Wrong Stuff," and "No Stuff." Authors don't say how many participants were in each cluster. Criterion measure was rate of promotion, which was operationalized as rank relative to self-reported total logged military flying hours. "Right stuff" participants had a significantly higher promotion rate than either of the other groups. No difference in promotion rate between "Wrong Stuff" and "No Stuff." There were also interpretable relationships between cluster membership and attitudes toward cockpit management as measured by the CMAQ.	Gregorich, Helmreich, Wilhelm, & Chidester (1989) (describe the 3 clusters)	Accdg to Musson et al., a number of the PCI scales had their origins in studies of gender differences in personality. Gender differences in a sample of 259 astronaut applicants (only 46 females) were not that large, with the exception of the Competitiveness scale on which males scored quite a bit higher (which is consistent with general population norms for this scale). Specifically, male astronaut applicants scored significantly higher on Competitiveness (p=.000) and Negative Instrumentality (p=.045) while Female astronaut applicants scored significantly higher on Expressivity (p=.040) and Achievement Strivings (p=.010). Authors note that the latter three findings would not be considered significant if Bonferroni corrections (based on # of comparisons) were made.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Personal Characteristics Inventory (PCI) **(Cont.)**				Howse (1995) collected data for 529 student aviators and 331 active duty aviators (no indication of gender breakdown)	Howse (1995)	PCI scales correlated significantly with NEO-FFI scales, in ways that make sense, but the correlations were mostly modest in size (~.25-.35) (N=147) (Musson, et al.) Musson et al note that the Positive Instrumentality/Expressive cluster ("Right Stuff") was related to higher scores on the NEO Conscientiousness factor. General literature shows that Conscientiousness is a predictor of performance in many (if not most or all) jobs (e.g., Mount & Barrick meta-analysis; Tett, et al. meta-analysis).
Cockpit Management Attitudes Battery; scales/clusters used in MTTB US Army	Personality: Cockpit Management Attitudes Questionnaire; Work & Family Orientation Scale; Revised Jenkins Activity Survey; Extended Personal Attributes Questionnaire	D: Rational, in that scales selected to measure specific constructs S: Likely some of each	yes, but not operationally	In a series of regressions, Intano & Howse report zero-order validities (among other things) for clusters of scores from the CMAB that were included in regression analyses. For predicting Primary Overall Flight Grade (PFOG), the regression included one cluster ("Cluster 1") that showed a zero-order correlation of .07 with PFOG (N=2405 pilot trainees). For predicting Primary Overall Average Grade (POAG), the regression included 3 non-cognitive clusters -- "Cluster 1," "Comp15," and "Cockpit" with zero-order correlations of .07, .10, and.02 respectively (N=2901 pilot trainees). Zero-order correlations were not reported for any non-cognitive scales that didn't make it into the regression analysis.	Intano & Howse (1991); Intano & Howse (1992); Intano, Howse, & Lofaro (1991a, 1991b)	cluster scores were created by combining scales from the different inventories; no description of how the clusters were created. Computerized, but old.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Cockpit Management Attitudes Questionnaire	3 underlying dimensions are Communication & Coordination; Command Responsibility; and Recognition of Stressor Effects				Gregorich, Helmreich, Wilhelm, & Chidester (1989)	This is an attitude instrument, not a personality measure. It was used as a criterion measure for a PCI study.
Automated Aircrew Personality Profiler USAF	Five factor solution extracted: hostility, self-confidence, values flexibility, depression, and mania 202 items: Mixture of types	D: Rational for picking inventories on which AAPP is based, but some of those were developed empirically S: Factor-analytic	yes, but not operationally	N=325 USAF UPT students. AAPP administered with the BAT. Hostility, self-confidence and values flexibility factor scores were significantly correlated with UPT pass/fail (-.12, -.13, -.12, respectively). (Authors note that > 80% passed, so upper limit on point-biserial was about .70). Entered the AFOQT subscales, BAT subtests, and the 5 AAPP factors scores in a regression equation to predict UPT graduate versus non-graduate. Sometimes included non-cognitive BAT subtests (Self-Crediting Word Knowledge & Activities Interest) and sometimes didn't. In all analyses, deleting the 5 AAPP factors scores reduced R, but never to a statistically significant degree so author concluded that the AAPP didn't add any unique prediction.	Siem (1992)	202 items from scales of MMPI, State-Trait Anxiety Inventory, Personal Orientation Inventory, Interpersonal Behavior Scale and Jenkins Activity Survey. Five factor solution extracted: hostility, self-confidence, values flexibility, depression, and mania. AAPP item pool may contain at least some of the response validity scales from the MMPI. Siem (1991) speculates that response latencies may be a viable method of dealing with faking, but his study doesn't directly address this question. Computerized.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Automated Aircrew Personality Profiler (Cont.)				N=332, same overall sample of USAF UPT students as reported in Siem (1992). Collected response latency for each item. Calculated 5 different scale scores based on the 94 MMPI items included in the AAPP (factor scales developed by Costa & colleagues in the MMPI item pool) -- Sociability (Psychoticism), Emotional Stability (Neuroticism), Extraversion, Competency (Inadequacy), and Cynicism. Three of the 5 factor scale scores showed low, but significant validity for predicting UPT P/F (Sociability .131, Competency .098, and Cynicism -.138). Factor analysis of the response latency and scale scores didn't produce a very interpretable factor solution. Conducted regression analyses to see if response latency added any predictive validity for UPT P/F. Basically, it didn't, with the possible exception of Extraversion, but even that finding seems pretty shaky. For example, the Extraversion scale score wasn't significantly correlated with UPT P/F to begin with.	Siem (1991); Siem (1992)	

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Pilot Personality Questionnaire US Navy	12 Personality scales: assertiveness, interpersonal orientation, aggressiveness, hostility, verbal aggressiveness, submissiveness, high-mastery motivation, high work motivation, competitiveness, self control, fatalism, high social desirability 112 multiple-choice items with no specific time limit	D: Rational, in that scales selected to measure specific constructs S: Unknown	yes, but not operationally	n=106 SNAs. Threw 69 predictor DVs in a regression to predict Advanced Flight Grade (AFG). The best regression solution included 23 of the DVs. Only 11 of the 23 showed a significant contribution ($p < .05$) to the regression, so further examined those 11. Three of the DVs were PPQ scales. Of the 3, only the Assertiveness scale showed a significant stand-alone correlation with AFG ($r=.20$, $p<.01$). The Assertiveness scale, when added to the AQT/FAR (which come from the ASTB) increased R-squared by .069 (R increased from .18 to .32). The Competitiveness scale, which didn't show a significant correlation with AFG as a stand-alone ($r=12$, ns), increased R-squared by .080 when added to AQT/FAR (R increased from .18 to .34). Another scale, Self-Control, which didn't show a significant correlation with AFG as a stand-alone ($r=.08$), increased R-squared by .04 (R increased from .18 to .27).	Street, Dolgin, and Helton (1993)	combination of Locus of Control, Work and Family Orientation, Personality Attributes Questionnaire, and Social Desireablity Scale. Includes the Crowne Social Desirability scale. This scale isn't a response validity scale, per se. Rather, it measures a person's tendencies to respond in a socially desirable manner which, in itself, can be interpreted as a personality construct. Computerized.
				N=45 SNAs; only the Crowne Social Desirability scale correlated significantly with UPT pass/fail (-.293) and flight grade (-.451), suggesting that those who scored lower on social desirability were more likely to pass and earn higher grades in UPT.	Shull & Dolgin, 1989	

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Pilot Personality Questionnaire (Cont.)				n=211 SNAs; compared those who passed all the way through advanced flight training with those who failed for academic or flight-related reasons (N=45). The PPQ Competitiveness scale was the only scale that differentiated those who passed from those who failed, in terms of mean score diff (t-test) and in a discriminant function analysis. SNAs with higher Competitiveness scale scores were more likely to pass.	Street, Helton, & Dolgin (1992)	
Tridimensional Personality Questionnaire US Navy (under review)	Personality: Novelty Seeking; Harm Avoidance; Reward Dependence 100 T-F items	D: Presumably rational S: Presumably rational	yes, but not operationally	to date, have only compared student naval aviators and flight officers with normative groups to check internal validity	Lambirth, Dolgin et al (2003)	test based on biosocial theory of personality. Computerized.
Hand US Navy (under review)	projective test that is scored according to 15 categories (listed below) & 6 summary scores, plus psychopathology index plus rigidity index Item type: Stimulus objects. 15 min. admin, 15 min to score	D: Unknown S: Clinical interpretation guided by scoring guidelines	yes, but not operationally	to date, have only compared student naval aviators and flight officers with normative groups to check internal validity	Lambirth, Dolgin et al (2003)	shown 10 cards, open-ended response to what the hand might be doing; presumably must be scored by a trained clinician. Virtually impossible to fake good on a projective test; may be possible to fake bad (or at least attempt to fake bad). Not Computerized.
	affection, dependence, communication, exhibition, direction, aggression, acquisition, active, passive, tension, crippled, fear, description, bizarre, failure					

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Jenkins Activity Survey US Navy examined	Composite Type A Behavior Pattern score, plus 3 factor-analytically derived subscales: Speed & Impatience (Factor S), Job Involvement (Factor J), and Hard-Driving & Competitive (Factor H); some articles say that scores for achievement striving and impatience/irritability can also be derived 20-30 minutes; 52 multiple-choice items	D: Unknown S: Factor-analytic, perhaps also empirical	yes, but not operationally	158 student naval aviators. No significant correlation between any Jenkins Scale and pass/fail or with grades in primary flight training.	Shull, Dolgin, and Gibb (1988)	Intended as a measure of Type A behavior, and for use as a diagnostic tool. Clinical research has shown that scores are predictive of coronary heart disease. Jenkins Activity Survey has been incorporated in numerous batteries of non-cognitive measures for research on pilot selection. Shull, Dolgin, & Gibb (1988) focused on Type A scale, and found that increased Type A characteristics did not lead to a higher probability of completing primary flight training; Review written by Center for Psychological Studies says the norm sample doesn't include women, young, elderly, or persons of low SES. Computerized.
NAMRL Automated Risk Assessment Task NAMRL (US Navy)		Unknown	Yes	440 student naval aviators. No correlation with any of the tests that make up the AQT/FAR. No correlation between any measure of risk taking and pass/fail or with grades in preflight or flight training.	Shull, Dolgin, and Gibb (1988)	Score is based on number correct and reaction time, so really not possible to fake good; might be possible to fake bad (I.e., intentionally get a low score). Computerized.
Defense Mechanism Test Norwegian/Swedish/Danish Air Force		Unknown	Yes	Meta analysis shows correlation of 0.3 to pass/fail for Scandinavians (n=1674) and 0.05 for other groups.	Martinussen and Torjuseen (1998)	Test of Freudian defense mechanisms. Used for selecting Norwegian AF pilots since 1978. Administered after acceptance into flight training. Not computerized.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Minnesota Multiphasic Personality Inventory Used widely by clinicians and private-sector organizations; may be used for psychiatric screening by the US military	10 clinical personality scales and 4 response validity scales (which are sometimes interpreted like content scales); numerous other scoring profiles have been empirically derived from the MMPI item pool 566 T-F items	D: Empirical S: Empirical	Yes	Profiles for various pilot samples have been developed and compared with general population norms; no evidence of validity for predicting training or job performance.	Caldwell, O'Hara, Caldwell, Stephens, & Krueger (1993)	The MMPI is most useful for psychiatric screening, I.e., "screen out" decisions. Is of limited use for "screen in" decisions. There are several response validity scales that are taken into account when scores are interpreted, including decision about whether or not scores can be interpreted at all. Probably Computerized.
Eysenck Personality Inventory UK for wide variety of jobs; UK military	Personality: Extraversion, Neuroticism 57 true-false items	D & S: Factor-analytic	No	495 selectees for British Army helicopter training. Score on extraversion scale correlated with pass/fail. No correlations with aptitude measures. Adding extraversion scale to aptitude battery score increases predictive validity about 4%.	Bartram and Dale (1982)	EPI probably doesn't measure enough of the qualities critical for helicopter pilot performance to be worth pursuing. Probably computerized.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Eysenck Personality Inventory (Cont.)				N=167 pilot trainees; criterion was successful completion of Basic Flight training versus suspension from same (N=58); Significant score difference between the two groups (success versus suspended) on the Neuroticism factor with suspended trainees scoring lower; no difference on Extraversion factor. Refined the results by categorizing each group member according to their quadrant on the 2x2 E-N factors, then calculating the percentage of failures falling in each quadrant. Found a significant chi-square primarily due to disproportionately high percentage of failures in the neurotic-introvert quadrant and disproportionately low percentage of failures in the stable-introvert quadrant. So, it's not as simple as avoiding neurotic candidates; should avoid neurotic introverts but consider selecting stable introverts.	Jessup & Jessup (1971)	
Inventories Developed and/or Used by Military for Some Purposes but not (yet) tried for Pilot Selection						

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Test of Adaptable Personality (TAP) US Army	80 Biodata items, but only 68 of them are scored	D: Rational S: Rational	No	N=31-48 US Army Special Forces Soldiers; four "rational biodata" scales (Intell Open, Tol of Ambiguity, Ach Orient, Fitness Motiv) were significantly correlated with peer ratings (.35-.45) and two (Ach Orient, Fitness Motiv) were significantly correlated with cadre ratings (external observer; .36 & .44) of performance in a 14-day field exercise; same study found lower (ns) correlations between Big 5 personality measures and Self-Efficacy measures with ratings of field exercise performance.	Summary documents provided by R. Killcullen, September, 2004; Kilcullen, Goodwin, Chen, Wisecarver, & Sanders (undated); Kilcullen, Mael, Goodwin, & Zazanis (1999)	scales can be added/dropped over time; several Kilcullen studies focus on "rational biodata scales" Kilcullen says these are TAP scales (personal communication). Use score on faking scale to adjust content scale scores. Not Computerized.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Test of Adaptable Personality (TAP) **(Cont.)**				N=210 US Army Special Forces solders; 20 "rational biodata" scales; calculated validity after excluding those who triggered more than 4 Lie scale items and keeping those who didn't trigger any of the Lie scale items (62% of 210). Criterion measures were supervisory ratings on BSS and self-reported awards & trg completed. Using the screened sample, Cognitive Flexibility, Work Motivation, & Ach. Orient showed significant correlations with supervisor ratings (.15, .22, & .25 respectively). Using the "Non-Faking only" sample, Work Motivation (.30), Ach. Orient (.37), Dominance (.22), Cognitive Flexibility (.18), Anxiety (-.17), and Fitness Motivation (.20) were significantly correlated with SF job performance. A backward stepwise regression based only on the rational biodata scales revealed that the combination of Work Motivation & Achievement Orientation yielded a significant prediction equation (no R reported). NOTE: This study actually found that cognitive, demographic, and fitness predictors didn't contribute to predictions of performance beyond the motivation variables (which were measured using rational biodata). However, the authors note that the sample was extremely range restricted on these dimensions (although, presumably, there was also at least indirect range restriction on the motivational variables).		Alphas for scales in Version 7 (Cog Flex, Fitness Mot, Peer Leadership, IPS-Team Player, IPS-Diplomacy, Work Motivation) range from .64 to .75 in a sample of 358 Army Special Forces candidates. These are fairly low, but not particularly surprising when you examine the items because some of the scales appear multidimensional.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Test of Adaptable Personality (TAP) (Cont.)	Work Motivation (aka Achievement Orientation)			r=.39 (peer rtgs of field perf); r=.36 (cadre ratings of field exercise perf); r=.25 or r=.30 for predicting supervisor ratings of field performance, in screened & non-faking samples, respectively.	Killcullen, Goodwin, et al.; Kilcullen, Mael, et al. (undated)	
	Cognitive Flexibility					
	Fitness Motivation			r=.45 (peer rtgs of field perf); r=.44 (cadre ratings of field exercise perf)	Killcullen, Goodwin, et al.	listed in all 3 Killcullen articles
	Peer Leadership					listed in Kilcullen "Summary of TAP"
	Team Orientation (aka Interpersonal Skills - Team Player)					listed in Kilcullen "Summary of TAP"
	Diplomacy (aka nterpersonal Skills - Diplomacy)					listed in Kilcullen "Summary of TAP"
	Faking (Unlikely Virtues)					listed in Kilcullen "Summary of TAP"
	Intellectual Openness			r=.37 (peer rtgs of field exercise perf); r=.15 (ns) (cadre rtgs of field exercise perf)		listed only in Kilcullen, Goodwin, et al.
	Tolerance of Ambiguity			r=.34 (peer rtgs of field exercise perf); r=.07 (ns) (cadre rtgs of field exercise perf)	Killcullen, Goodwin, et al.	listed only in Kilcullen, Goodwin, et al.; scale name suggests it MAY be the same as Cognitive Flexibility
	Affective Commitment to the Army					"new" scale being added, accdg to Kilcullen "Summary of TAP"
	Stress Tolerance					"new" scale being added, accdg to Kilcullen "Summary of TAP"

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Test of Adaptable Personality (TAP) (Cont.)	Narcissism					"new" scale being added, accdg to Kilcullen "Summary of TAP"
	Hostility to Authority					Scale taken from Assessment of Right Conduct tool; has been developed and validated against various delinquency criteria (Kilcullen, personal communication, February 7, 2005)
	Respect for Authority					experimental 4-item scale has already been written
	Internal Locus of Control					"new" scale being added, accdg to Kilcullen "Summary of TAP"
	Self-Efficacy					"new" scale being added, accdg to Kilcullen "Summary of TAP"
	Cultural Tolerance					"new" scale being added, accdg to Kilcullen "Summary of TAP"
	Attentional Focus					"new" scale being added, accdg to Kilcullen (personal communication, February 2005)

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Assessment of Individual Motivation (AIM) US Army	Personality: Dominance (Leadership); Work Orientation; Agreeableness; Dependability (Nondelinquency); Adjustment; Lie Scale 30 minutes	D: Rational S: Rational (forced choice)	No	White, Young, & Rumsey (2001) report that the AIM predicted attrition among Army enlisted personnel (N=5000+; r=-.15) and showed a similar level of validity for predicting attrition in a sample of 8,500 USAF enlisted personnel. These say that the AIM also showed validity for predicting performance and personal discipline among correctional officers in military prisons and success in explosive ordinance disposal training for military personnel.	White, Young, & Rumsey (2001); White & Young (1998); White, Young, Heggestad, Stark, Drasgow, & Piskator, 2004, 2005)	The AIM is a forced-choice inventory designed to measure the key constructs measured by the ABLE. The scales show considerable convergence with the ABLE (by design) but less susceptibility to faking. Can correct scores or screen on basis of Lie scale, although this isn't currently done operationally. Computerized.
						AIM is used operationally for pre-enlistment screening of non-High School Graduate recruits and for some in-service testing. It is also part of a test battery being evaluated for selection of non-commissioned officers (Part II of the Non-commissioned Leadership Skills Inventory).
						Currently, items are presented in tetrads, but the Army Research Institute is working on a paired comparison version.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
Assessment of Background & Life Experiences (ABLE) US Army	Achievement Orientation; Dominance (Leadership); Dependability; Adjustment; Cooperativeness; Internal Control; Physical Condition; Social Desirability; Nonrandom Response 70-199 Personality & Biodata items depending on which version is selected & which resp val scales included	D: Rational S: Rational	No	Project A research showed that the ABLE was moderately valid for predicting "will-do" aspects of enlisted Soldier performance. It also showed incremental validity beyond cognitive measures for predicting will-do aspects of job performance.	White, Young, & Rumsey (2001)	Army researchers have developed several versions of the ABLE, including versions that use fewer items to measure the same constructs (ABLE-114, ABLE-70); a forced-choice version (AIM); and two different biodata versions: 1 incorporated personal history items but was not constrained to the use of objectively varifiable items, the other was so constrained (latter were called biodata analog scales). Can screen on the basis of response validity scale scores Not Computerized.
				Scale profile has been developed for aviator sample, could potentially be used for profile matching; however, visual inspection of scale profiles for a student versus an aviator sample suggest that the profiles are pretty similar; quantitative analysis may suggest some differences.	Howse (1995)	
Assessment of Right Conduct (ARC) US Army	Social Maturity; Self-Esteem; Traditional Values; Manipulativeness; Greed; Hostility to Authority; Response Distortion. 80 Biodata items, ~ 20 minutes	D: Rational S: Rational	No	construct validity evidence available, but no criterion-related validity evidence (yet)	Kilcullen, White, Sanders, & Hazlett (2003)	Has been used with Special Forces Soldiers Use score on faking scale to adjust content scale scores Not Computerized.

Name	Scales	Development & Scoring Approach	Been Tried for Pilot Selection?	Summary of Validity Evidence	Key Reference(s) Describing Test or Documenting Validity	Comments
California Psychological Inventory (CPI) used widely by clinicians and private-sector organizations	20 scales, plus 3 "vectors" (Internality, Norm-Favoring, & Self-Realization); several other scoring profiles have been developed over the years. 434 Personality items, ~ 45 minutes	D: Some scales were empirically derived; others rationally S: Empirical	No		Gough & Bradley (1996); Howse (1995)	The CPI is primarily useful as a clinical assessment tool. It has been used for personnel selection purposes, but generally is best suited for situations in which a clinical intepretation of the score profile can be utilized. Scale profile has been developed for aviator sample, could potentially be used for profile matching; however, visual inspection of scale profiles for a student versus an aviator sample suggest that the profiles are fairly similar; quantitative analysis may suggest some differences.

Appendix C
Recommended Selection Strategy for Army Aviators

Broad Construct that Should be Measured	Recommended Stage	Best Bet Predictor Measure	Approx Admin Time	Pros & Cons
Cognitive Ability - Spatial - Quantitative - Verbal - Mechanical (ASTB) - Perc. Speed & Acc (AFOQT)	1	All cognitive subtests from the: 1. ASTB (US Navy) or 2. AFOQT (USAF)	3 - 4 hours	Both the ASTB and the AFOQT have a proven track record for aviator selection, and either should be available to the Army for no charge. The ASTB has the decided advantage of already being programmed for administration via the Internet. The AFOQT includes more narrowly-defined set of subtests than the ASTB, but both cover the broad domain of cognitive ability. We evaluated the difficulty of the both test batteries and found no evidence that either one would be too difficult, overall, for the Army Aviator applicant population.
Perceptual Speed & Accuracy	1	Table Reading Test OR Publicly-available measure OR Newly-developed measure	part of AFOQT admin time if that battery is selected; ~10-15 minutes if administered as a stand-alone test	The Table Reading test, which is a measure of Perceptual Speed & Accuracy, has proven validity for predicting aviator performance, it requires only a small amount of time to administer, and adds unique variance to the prediction of aviator performance. It is part of the AFOQT, and has been since the 1940's. If the Army does not want to administer the AFOQT in the preliminary validation study, but does want to measure this important skill, it could administer a commercially-available version of the Table Reading Test that is owned by DAS. It could also explore other measures of Perceptual Speed & Accuracy that are available in the public domain or ask PDRI to develop a new measure.

Broad Construct that Should be Measured	Recommended Stage	Best Bet Predictor Measure	Approx Admin Time	Pros & Cons
Motivation / Attitude	1	Army-specific measure of Helicopter/Aviation Information OR Aviation & Nautical Information (ANI) subtest from the ASTB OR Aviation Information (AI) subtest from the AFOQT	20-30 minutes [Admin time for ANI or AI subtest is included in the admin time for the ASTB or AFOQT]	PDRI could, with assistance from Army SMEs, develop content for an Army-specific Helicopter/Aviation Information subtest. If the Army decides to administer the ASTB or the AFOQT in the preliminary validation study, the ANI or the AI could also be administered. Neither the ANI nor the AI has high face validity for an Army aviation applicant sample, but both have shown relatively high empirical validity for predicting performance during aviator training, probably because they act as a measure of motivation to become an aviator. It's reasonable to assume that these subtests will serve as a measure of motivation for Army aviation applicants too, and we could gather empirical data necessary to test this assumption. Of the two Information subtests, the ANI has the decided advantage of already being programmed for administration via the Internet.
Personality/Temperament	1	Selected scales from the US Army's Test of Adaptable Personality (TAP) or Assessment of Individual Motivation (AIM), USAF's Self-Description Inventory, and/or USAF's Armstrong Laboratory Aviation Personality Inventory (ALAPS) And/or New, highly-tailored scales developed by PDRI	~90 minutes for validation study; ~45-60 for operational version	No single inventory measures all of the characteristics likely to be important for the aviator job but, across the four inventories listed here, there is good coverage of many important traits. When the job analysis results are available, we will select a set of non-redundant scales from these three inventories and/or write a small number of new scales that measure the most critical personality/ temperament characteristics. In the newly-configured inventory, some items will be biodata-like, others will be personality-like. All four of the existing inventories belong to the US military so there should be no charge to use any of the scales. We can't use the ALAPS without modification because the items and scoring were published in a USAF technical report.

Broad Construct that Should be Measured	Recommended Stage	Best Bet Predictor Measure	Approx Admin Time	Pros & Cons
Cognitive Task Prioritization	1 if Popcorn Test; 2 if WOMBAT	Popcorn Test or WOMBAT	~40 minutes [included in the WOMBAT admin time if that battery is used]	Cognitive task prioritization has been measured for many years by NASA using a Popcorn type test, although this type of test has never been used for selection purposes. Popcorn tests are not difficult to develop and DAS, with programming assistance from AIR, could develop one pretty quickly for relatively little cost. The WOMBAT also measures this ability, but it may not be used at all and, even if it is, will likely not be feasible as a Stage 1 selection tool, so if the Army wants to include a measure of this important ability in the Stage 1 screen, it should develop a Popcorn test.
Psychomotor 1. Multi-Limb Coord. 2. Tracking	2	1. TBAS (USAF and US Navy) 2. WOMBAT 3. newly-developed psychomotor test battery	60-90 minutes	The TBAS is computerized (but not for web administration), showed incremental validity beyond the AFOQT in a small study, and is free. However, preliminary examination raised concern about the programming and we've been unable to locate any programming documentation or information about how dependent score variables are derived. WOMBAT is a commercially-available test used extensively in Canada for aviator selection, although there isn't much published validity evidence. It's a 90-minute test that can be shortened by reducing the testing period. The drawback to using a shorter version is that the test no longer measures persistence on a cognitively-taxing task. The developers are working on a web-administered version. Another alternative is to ask our team (DAS/AIR) to develop a psychomotor battery. The battery could measure psychomotor skills as well as TBAS (with better documentation) but wouldn't be as comprehensive or complex as WOMBAT. Time, budget, and ownership issues would have to be negotiated.

Broad Construct that Should be Measured	Recommended Stage	Best Bet Predictor Measure	Approx Admin Time	Pros & Cons
Multi-Task Performance	2	1. TBAS (USAF and US Navy) 2. WOMBAT 3. newly-developed test	included in the 60-90 minute admin time listed above	same as above; If DAS/AIR developed a new test, they would combine a secondary task with the psychomotor test battery to derive a measure of multi-task performance
Situational Awareness/ Stress Tolerance	2	WOMBAT	included in the 60-90 minute admin time listed above	The WOMBAT arguably measures situational awareness and stress tolerance, both of which are important skills for aviators. Gaining measures of these abilities would be an advantage of administering the WOMBAT. However, a shortened version clearly would not measure stress tolerance as well as the original 90-minute version.