U.S. Air Force Pilot Selection Tests: What is Measured and What is Predictive?

THOMAS R. CARRETTA, M.S., Ph.D., and MALCOLM JAMES REE, M.A., Ph.D.

THE AIR FORCE OFFICER QUALIFYING TEST (AFOQT, 6) and the Basic Attributes Test (BAT, 5) are important components in U. S. Air Force (USAF) pilot selection (2). These tests have been described as measuring a variety of abilities that have demonstrated utility in predicting pilot success (9). Recent studies using the AFOQT and BAT suggest that their predictive utility, beyond g, of measures of pilot job knowledge (e.g., aviation concepts, instruments, principles and terms; increase in correlation beyond g of about 0.08), psychomotor abilities, and personality scores (increase in correlation beyond g between 0.02 and 0.04) also has been small, but significant. The unavoidable requirement to reason in responding to test material causes g to be measured. In broad-ability-range samples, the positive correlations of the measures demonstrate that general cognitive ability is always present as a higher-order factor. Future measures of pilot aptitude may include tests based on cognitive components, chronometric methods, neural conductive velocity, or other methods. These measures, despite their appearance, have been shown to mostly measure g. Subsequently, we expect that future U. S. Air Force pilot selection tests will mostly be measures of g and will, therefore, continue to be predictive of performance.

Validation Strategies

"Validity" is the term used to signify that an independent variable (e.g., ability test, interview, selection composite) predicts a dependent variable (e.g., training performance, job performance, promotion, accidents). To determine if a test or any other measure is useful for prediction, a "validation study" is conducted. In this study, a criterion is regressed on a predictor or set of predictors. In general, there are two types of empirical validation studies: predictive and concurrent. In the predictive study unscreened applicants are tested prior to selection. Criterion data are collected after the subset of those selected complete training or are rated on some aspect of job performance. In the concurrent study, those already selected or those on the job are tested. Criterion data are collected shortly thereafter.

"Incremental validity" refers to the ability of one predictor to add to the predictive efficiency of another predictor. For example, if a measure of intelligence predicts job performance at 0.50 and the measure of intelligence combined with a measure of psychomotor ability predicts performance for the same sample and criterion at 0.65, the difference between 0.50 and 0.65 (0.15) is the incremental validity of the psychomotor measure.

Criteria for validation studies can range from training to tenure to number of accidents. In the studies reviewed here we frequently used criteria derived from training. This is because job performance criteria are very expensive to collect and may be disruptive of normal operations. In one study, job performance criteria were used.

Training criteria are important as training is a prerequisite to job performance. Training imparts job knowledge that facilitates the performance of the job tasks. Several criteria exist in the 53-week USAF pilot training program. The first is the dichotomous pass/fail training. Others include classroom grades for such subjects as aircraft instruments, navigation, flight physiology, and mission planning. Further criteria are a series of daily flights and check flights that are work samples of flying performance. To avoid familiarity bias, these work samples are rated by check flight pilots who did not fly with student
pilots. There are also flying training rankings that combine the classroom grades, daily flights, and check flight. In some studies, additional composites of overall performance were derived by summing several criteria. Finally, supervisory and peer ratings of on-the-job performance were also used. The performance rating factors were derived from job analyses and are comprised of ratings of tasks such as "developing and executing a tactical game plan," "weapons systems proficiency," and "analyzing engagement geometry."

Now that an overview of validation has been discussed, the next section will review current U.S. Air Force pilot selection tests. Their content, factors, and sources of validity are presented for several pilot performance criteria.

**Currently Used USAF Pilot Selection Tests**

**Air Force Officer Qualifying Test**

The Air Force Officer Qualifying Test (AFOQT) has been used by the Officer Training School and the Reserve Officer Training Corps since 1957 for officer commissioning and aircrew training selection (6,29). The 16 AFOQT subtests are combined into 5 operational composites (Verbal, Quantitative, Academic Aptitude, Pilot, and Navigator-Technical; 29). Skinner and Ree (29) reported that the AFOQT measured five lower-order factors (i.e., verbal, quantitative, spatial, aviation interest/aptitude, and perceptual speed) that closely mirrored the five composites used in USAF officer and aircrew selection. They did not, however, investigate the hierarchical factor structure of the AFOQT, nor confirm their analyses with statistical tests. Recent confirmatory factor analyses (7) have demonstrated that the AFOQT has a hierarchical nature similar to other multiple aptitude tests (14,22). In these analyses, the average correlation among the 16 AFOQT subtests was about 0.44 and the proportion of common (i.e., explained) variance due to general cognitive ability was estimated to be 67%. The remaining common variance (33%) was distributed among the residualized lower-order factors of verbal (11%), quantitative (4%), spatial (4%), aviation interest/aptitude (9%), and perceptual speed (4%). The proportions for g and specific factor variances found in the AFOQT were similar to those found in other multiple aptitude batteries (11).

Ree and Earles (26,27) have shown that for cognitive test batteries, several methods may be used to estimate g and specific abilities (s) for participants for use in validation or other applied studies. These methods are all variants of factor analysis. Unrotated principal components and principal factors, respectively, factor all the variance or only the common variance in a matrix of correlations. Rotation is avoided in order to keep from distributing the first factor variance over the succeeding factors. In a hierarchical factor analysis, a matrix of correlated factors is further factored to reveal "higher-order" factors. All three of these methods, unrotated principal components, principal factors, and hierarchical factors, yield estimates of g and s that are uncorrelated with each other. This condition is mandatory for conducting studies of incremental validity. Ree and Earles (26,27) showed that these three methods produce estimates of g that correlate with one another at 0.98 or greater.

Carretta and Ree (6) examined the validity of the 16 AFOQT subtests against five USAF pilot training criteria based on academic grades, performance on daily (i.e., training) flights, and check (i.e., test) flights. The rank correlation of the general cognitive ability factor loadings of the AFOQT subtests with the average validity of the subtests for predicting these five pilot training criteria was 0.62. The higher the g loading of the subtest, the more valid it was for predicting the criteria.

Olea and Ree (20) conducted a comparison of the validity of general cognitive ability, g, and specific abilities and specific knowledge, s1...sn, for predicting pilot criteria in samples ranging from about 1800-3900 student pilots. General cognitive ability and specific abilities and knowledge were estimated as the first (g) and succeeding (s1...sn) principal components from the AFOQT. The criteria included an overall performance composite, academic performance, and work samples of landings, loops, and rolls for pilots attending training. An overall performance composite was created by summing academic performance grades and the work sample ratings. Regression equations were compared to evaluate the predictive efficiency of g and s for each of the criteria. The measure of g was the best predictor of each of the criteria, while s contributed little. The average of the validity coefficients for g across all pilot criteria was about 0.31, while the average of the incremental validities for the specific abilities and knowledge was about 0.098, with no overlap in the two distributions of correlations. The validity of g was 0.209-0.430 vs. 0.075-0.139 for the validity of s1 to sn. Further, little incremental validity was found for the composite performance criteria for pilots (0.089). The same was true for the work sample criteria. Results suggested that the incremental validity of specific measures was due to specific knowledge about aviation (i.e., aviation principles, controls, and instruments) rather than specific cognitive abilities (i.e., verbal, quantitative, spatial, or perceptual speed).

Ree, Carretta, and Teachout (25) investigated a causal model of the influence of general cognitive ability and prior pilot job knowledge (both from AFOQT subtests) on pilot job knowledge acquired during training and flying training work sample performance. They found that general cognitive ability exerted its influence on pilot training performance mostly through its influence on the acquisition of job knowledge. Prior job knowledge showed a causal effect on job knowledge acquired during training. Prior job knowledge also showed a causal influence on flying work sample performance early in training. As with previous research (20), the measure of g showed a much stronger causal influence on training performance than did the prior job knowledge.

**Basic Attributes Test and the Pilot Candidate Selection Method**

The Basic Attributes Test (BAT; 1, 2, 5) is another important component in USAF pilot selection. It includes...
measures of information processing speed, psychomotor performance, and personality (attitudes toward risk). The Pilot Candidate Selection Method (PCSM; 2, 5) combines scores from the AFOQT and BAT along with a measure of previous flying experience to create a pilot selection composite. PCSM has been used operationally by the Air National Guard as a component of their pilot selection procedures for over 5 yr, and is currently under review by the Air Force.

The PCSM model (2) combines the AFOQT Pilot composite with BAT psychomotor, cognitive, and personality measures, and a self-report of flying experience in hours flown. The psychomotor scores are combined in a unit-weighted composite (converted to standard scores and then summed with an implicit weight of 1). The psychomotor and AFOQT Pilot composite scores are regression-weighted along with the other PCSM components (cognitive, personality, flying experience) to predict pilot training performance.

An evaluation of the validity and incremental validity of the PCSM components was reported by Carretta and Ree (5) for a sample of 676 pilot trainees. Analyses of the validity of the PCSM components showed the following observed correlations with pass-fail pilot training: AFOQT Pilot composite 0.17, BAT psychomotor scores 0.15, information processing 0.06, personality 0.10, and previous flying experience 0.17. Adding all the scores collected on the BAT to the AFOQT Pilot composite incremented the correlation from 0.17 to 0.30.

These validity and incremental validity correlations show the effect of the phenomenon known as "range restriction." Thorndike (31) in discussing pilot selection during WWII presented a classic example. He reported data from a study in which 1036 men were admitted to pilot training without regard to scores on a selection battery. He then computed the validity correlations of this battery against training performance in two ways. First, the validities were computed in the sample of all men entering training (n = 1036); then they were computed in a subsample (n = 136) that would have qualified if existing standards had been used for pilot training. The correlations computed in the sample of those who would have "passed" showed the effects of range restriction. The correlations decreased on average from 0.41 in the unrestricted sample to 0.13 in the restricted sample. This is not an isolated phenomenon. It happens quite regularly and uniformly when samples that have been selected on the basis of one or more predictors are analyzed. However, there are well-known methods for correcting these statistical biases (correction for range restriction), and we have routinely applied them (31).

The correlations reported by Carretta and Ree (5) could not be corrected for range restriction due to linear dependence between the Pilot composite and the subtests of the AFOQT. These uncorrected correlations should be considered underestimates of the population values. The way to alleviate the problem of inability to correct for range restriction is to use the AFOQT subtests instead of the Pilot composite, thus avoiding linear dependency.

The incremental validity of the BAT scores and previous flying experience relative to the AFOQT subtests was estimated using correlations corrected for range restriction (5). To provide for best estimates of the correlations and to remove linear dependencies, the scores included the AFOQT subtests, BAT psychomotor, cognitive, and personality subtests, and previous flying experience. Regression equations were computed and compared to examine the validity and incremental validity of the scores. The AFOQT subtests, which have been shown to be highly g-loaded (7,8), were the best predictors of pilot training performance with a multiple correlation of 0.31. A multiple R of 0.38 was found when the AFOQT subtests, BAT scores, and previous flying experience were combined in the same regression equation. This increment of 0.07, an increase of 23% above the AFOQT subtests alone, generates large cost avoidance savings for the Air Force. The size of the increment for the BAT (which is a measure of g, psychomotor skills, and personality) and for previous flying experience (which measures, at least in part, pilot job knowledge), was consistent with the findings of Olea and Ree (20) who demonstrated that the validity of the AFOQT subtests was due to their measurement of g and pilot job knowledge.

Analyses were conducted to discover the causes of the amount of the incremental validity of the various predictors. The BAT psychomotor scores were studied in the presence of a highly g-loaded battery of verbal and mathematical cognitive tests (23). Confirmatory factor analyses on a sample of 354 enlisted personnel showed a hierarchical structure including seven factors. There were two lower-order cognitive factors representing verbal and math, and three lower-order psychomotor factors representing the BAT psychomotor tests of Two-Hand Coordination, Complex Coordination, and Time Sharing. A higher-order psychomotor factor influencing all psychomotor scores also was found. Unexpectedly, a higher-order general cognitive factor was found to influence all scores, both cognitive and psychomotor.

The general cognitive factor, g, accounted for 39% of the common variance and the general psychomotor factor accounted for 29%. The proportion of common variance accounted for by the lower-order factors was 10% for Two-Hand Coordination, 7% for Complex Coordination, 7% for Time Sharing, 5% for verbal, and 3% for math. This demonstrated that the BAT psychomotor tests measured, at least in some part, general cognitive ability. This is consistent with past research (10,12,30) showing that g can be measured in many different ways.

The finding that the psychomotor scores measured g as well as psychomotor skills suggested why they did not demonstrate greater incremental validity in the prediction of pilot criteria (5). There is limited reliable variance in the psychomotor scores that is unique with respect to cognitive ability. This limits their potential for incremental validity.

Ree and Carretta (23) speculated that the correlations between cognitive and psychomotor tests and the existence of a general cognitive ability factor that included both cognitive and psychomotor scores was due to the requirement to reason while performing the tests. The finding that the BAT psychomotor scores measured g is consistent with Rabbitt, Banerji, and Szymanski (21). They observed that "Space Fortress," a test that appears to require complex psychomotor-perceptual ability, correlated with paper-and-pencil IQ tests about as well as paper-and-pencil IQ tests correlated with each other.
PILOT SELECTION TESTS—CARRETTA & REE

Again, the need to reason and learn to apply complex rules may account for the findings of Rabbitt et al. (21).

The finding of small incremental validity for the personality variables used in PCSM (attitudes toward risk) is similar to that of McHenry, et al. (18). In a series of regression analyses, they showed that general cognitive ability was the best predictor of core job performance and was incremented by only about 0.02 when temperament and personality measures were added. However, when the performance criteria consisted of ratings more closely associated with personality (e.g., effort, leadership, personal discipline, and physical fitness), personality measures demonstrated larger increments to the predictiveness of g. The implication is that personality variables would show greater incremental validity if pilot criteria were influenced more by personality.

That the BAT tests based on information processing speed were not predictive was surprising (5) as they had been found to be predictive of pilot training performance in other studies (1,2). Their lack of validity may have been the result of sampling error. That they were not incrementally valid was not surprising, because of their degree of g-saturation.

Future Trends

The USAF is conducting several lines of research in the measurement of ability that may contribute to future pilot selection instruments. They include tests based on cognitive components, chronometric measures, and neural conductive velocity.

Cognitive components are often measured in computer-based tests. In a typical cognitive components measurement paradigm, participants may be shown a series of letters and sequentially presented a set of rules governing the ordering of the letters. Then, following the rules, participants must state the proper order of the letters.

Measures based on cognitive components such as information processing speed and working memory capacity have been shown to be mostly measures of general cognitive ability (15,17,19). Kyllonen (16) described several tests based on the cognitive components of declarative knowledge, procedural knowledge, processing speed, and working memory. In a previously unpublished study, the current authors reanalyzed Kyllonen's (16) data after correcting the observed correlations for range restriction. A hierarchical factor representing general cognitive ability was extracted and found to account for about 79% of the common variance among these tests. These cognitive-components-based tests were more g-saturated than either the AFOQT or BAT.

The USAF recently completed two studies predicting pilot performance that used experimental tests based on cognitive components (3,4). Results have shown several of them to be predictive of pilot performance. However, the early state of development of these new tests precludes comparisons of their validities with the validities of operational tests.

Examples of chronometric measures include tests based on reaction time, choice reaction time, and neural conductive velocity. Simple reaction time tasks that involve the ability to give one fast response to one signal (light, picture, sound, etc.) when it appears, have shown low positive correlations with measures of g (11,13). Choice reaction time tasks that require the ability to choose from among several potential alternatives and make an appropriate response (e.g., pressing the one lighted button among eight) have shown moderate correlations with measures of g (13).

Neural conductive velocity is the speed at which a neuron transmits an impulse. It requires no physically invasive procedure and is typically measured in the optic nerve. Electrodes are attached to the head and a light is flashed that the participant sees. No overt response is required by the participant. The verbal instructions are as simple as “look at the light.” Reed and Jensen (28) have shown that the neural conductive velocity along the optic nerve is correlated about 0.37 with measures of general cognitive ability.

Approaches based on cognitive components, chronometric measures, and neural conductive velocity are considered promising sources for new pilot selection tests because they measure mostly general cognitive ability. They offer the advantage of being content-free, thereby potentially reducing mean test score differences that were the consequences of differential educational choices.

Summary

The measurement of g is inherently unavoidable in all measures of ability (24). In broad-ability-range samples, the positive correlations of the tests demonstrate that general cognitive ability is always present as a higher-order factor. The unavoidable requirement to reason in responding to test material causes g to be measured regardless of whether the test requires a psychomotor response, specialized knowledge, or verbal skills. General cognitive ability is and should be measured with a wide variety of test contents (14). No single test can measure g as well as a widely varied battery.

A review of several recent studies involving the use of USAF pilot selection tests demonstrated that despite differences in appearance and stated test content, most are to a large extent measures of general cognitive ability or g. These tests range from paper-and-pencil aptitude batteries such as the AFOQT to the psychomotor portions of the BAT to computer-based measures of cognitive components. Further, the validity of these tests against pilot training criteria comes mostly from their measurement of g. The incremental validity beyond g of specific cognitive abilities such as verbal, quantitative, spatial, or perceptual speed has been shown to be small or nonexistent. Likewise, the incremental validity of pilot job knowledge (e.g., knowledge of aircraft concepts, instruments, principles, and terms), psychomotor abilities, and personality scores has been shown to be small, but significant.

Future measurement of pilot aptitude may include tests based on cognitive components, chronometric methods, or neural conductive velocity. These types of measures also have been shown to be highly g-loaded. Subsequently, it is expected that future USAF pilot selection measures will continue to be mostly measures of general cognitive ability and will therefore, continue to be useful in the prediction of pilot performance.
ACKNOWLEDGMENTS
The views expressed are those of the authors and are not necessarily those of the U.S. Air Force, Department of Defense, or the U.S. government.

REFERENCES