AFRL-SA-WP-TR-2012-0004

PILOT COGNITIVE FUNCTIONING AND TRAINING OUTCOMES

Raymond E. King
U.S Air Force School of Aerospace Medicine

Paul Retzlaff
Eagle Applied Sciences, San Antonio, TX

Erica Barto, Malcolm James Ree
Operational Technologies Corp., San Antonio, TX

Mark S. Teachout
University of the Incarnate Word

Thomas R. Carretta
Air Force Research Laboratory

July 2012

Final Report
for September 2011 – July 2012

Distribution A: Approved for public release; distribution is unlimited.
Case Number: 88ABW-2012-4567,
21 Aug 2012

Air Force Research Laboratory
711th Human Performance Wing
School of Aerospace Medicine
Aeromedical Research Department
2510 Fifth St.
Wright-Patterson AFB, OH 45433-7913

NOTICE AND SIGNATURE PAGE

AFRL-SA-WP-TR-2012-0004 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//SIGNATURE//	//SIGNATURE//
Dr. William W. Dodson, Chief/FHC	Col Michael D. Grinkemeyer, Chair/FH

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
31 Jul 2012

2. REPORT TYPE
Final Technical Report

3. DATES COVERED *(From – To)*
September 2011 – July 2012

4. TITLE AND SUBTITLE

Pilot Cognitive Functioning and Training Outcomes

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Raymond E. King, Paul Retzlaff, Erica Barto, Malcolm James Ree, Mark S. Teachout, Thomas R. Carretta

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
USAF School of Aerospace Medicine
Aeromedical Research/FHC
2510 Fifth St.
Wright-Patterson AFB, OH 45433-7913

8. PERFORMING ORGANIZATION REPORT NUMBER

AFRL-SA-WP-TR-2012-0004

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A: Approved for public release; distribution is unlimited. Case Number: 88ABW-2012-4567, 21 Aug 2012

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The current study examined the relationship between clinical cognitive functioning tests and U.S. Air Force pilot training outcomes. Three computerized tests were used: the Multidimensional Aptitude Battery, MicroCog, and CogScreen. In addition to the traditional pass/fail training outcome, the quality of passing and reasons for failure were examined. Outcome criteria for training graduates included class rank, academic grades, daily flying grades, and check ride grades. Reasons for failure included flying training deficiency and being "Dropped on Request." Correlations in samples of between 5,582 and 12,924 trainees across the tests showed small, but important, relationships with training outcomes. All three of the clinical tests performed similarly. There was little evidence that any specific cognitive variable was more important than any other, and the results pointed to general cognitive ability as the main predictor of performance. In terms of the outcome variables, performance for graduates (e.g., class rank) was better predicted than training attrition.

15. SUBJECT TERMS
Neuropsychological testing, aviators, cognitive ability, training outcome

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Lt Col Raymond E. King
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	SAR	37	

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

This page intentionally left blank.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

1.0 SUMMARY

The current study examined the relationship between clinical cognitive functioning tests and U.S. Air Force pilot training outcomes. Three computerized tests were used: the Multidimensional Aptitude Battery, MicroCog, and CogScreen. In addition to the traditional pass/fail training outcome, the quality of passing and reasons for failure were examined. Outcome criteria for training graduates included class rank, academic grades, daily flying grades, and check ride grades. Reasons for failure included flying training deficiency and being "Dropped on Request" (DOR). Correlations in samples of between 5,582 and 12,924 trainees across the tests showed small, but important, relationships with training outcomes. All three of the clinical tests performed similarly. There was little evidence that any specific cognitive variable was more important than any other, and the results pointed to general cognitive ability as the main predictor of performance. In terms of the outcome variables, performance for graduates (e.g., class rank) was better predicted than training attrition.

2.0 INTRODUCTION

The U.S. military services and the U.S. Air Force (USAF) in particular have long histories of studying the selection of candidates for pilot training. Military aviation is an exceptionally demanding profession. The training of military aviators is long, difficult, and extremely expensive. While the majority of aviation candidates are successful in training, the cost of those who fail is a lost investment. In an attempt to screen candidates with fairness, reliability, and efficiency, psychological tests usually have been used.

A very comprehensive review of aviation testing and selection was commissioned by the U.S. Army and accomplished by Paullin, Katz, Bruskiewicz, Houston, and Damos (Ref 1). Here, cognitive as well as personality testing was reviewed with an eye toward the selection of pilot training candidates. They concluded that selection should follow the lead of the U.S. Navy and USAF in the use of those services' selection tests. They suggest that the Army look at using the Aviator Selection Test Battery, the U.S. Navy's primary aviator selection instrument (Ref 2). They also suggest the use of the Air Force Officer Qualifying Test (AFOQT) (Ref 3). Both tests were recommended due to their emphasis on assessing intelligence, cognitive ability, and information processing.

In the past year, Howse and Damos (Ref 4) have updated that work with a very comprehensive, 275-page annotated bibliography. The work was published through the Air Force Personnel Center. Of particular note is that two DVDs are available from the Air Force Personnel Center /DSYX that contain not only the references but all of the digital files associated with the project. This compendium is referred to as the "Digital Library of the History of Pilot Training Selection." Interested readers are referred to the publication, and researchers to the archive.

Throughout these reviews and other work, it appears that intelligence and cognitive functioning are key to successful pilot training completion. Indeed, Carretta and Ree (Ref 3) specifically suggested that general intelligence is by far the largest factor in the determination of pilot training success. Additional predictors include aviation knowledge, psychomotor ability, and, perhaps, personality.

2.1 Pilot Candidate Selection

Becoming a USAF pilot first requires being accepted into training, which is a long and arduous process. Initial pilot selection is accomplished through two basic methods. Regardless of commissioning source, all applicants must pass the rigorous Class I flight physical standards to be eligible for selection. Then, each commissioning source considers measures of aptitude and officership. USAF Academy cadets are selected by Academy faculty and staff, who take into account academic [e.g., grade point average (GPA)], physical, and military performance. Applicants who are commissioned through the Reserve Officer Training Corps (ROTC) or Officer Training School (OTS), including the Airman Education and Commissioning Program, are administered the AFOQT (Ref 5) and Test of Basic Aviation Skills (Ref 6). The AFOQT Pilot composite and several Test of Basic Aviation Skills subtest scores are combined with flying experience in a regression-weighted equation to create a measure of pilot training aptitude called the Pilot Candidate Selection Method. For ROTC, medically qualified pilot training applicants are ranked on their Order of Merit scores. This score is based on the Pilot Candidate Selection Method score, field training, physical fitness, college GPA, and commander’s ranking. OTS selection is based on the “whole person” concept, where applicants receive points over three areas: experience/leadership, education/aptitude, and potential/adaptability. A theme throughout all of these selection procedures is high intelligence, whether it involves being accepted into the Air Force Academy, a high GPA, a high AFOQT score, or the impression a candidate makes on a selection board member.

2.1.1 AFOQT. The most explicit test of cognitive ability and intelligence is the Air Force Officer Qualifying Test. The AFOQT is a paper-and-pencil multiple aptitude battery used for officer commissioning and aircrew training selection (Ref 7). It was developed and is maintained by the USAF. Administration time is about 3.5 hours. The 11 AFOQT subtests are combined to create 5 operational composites: Verbal, Quantitative, Academic Aptitude, Pilot, and Navigator-Technical. It has a hierarchical factor structure and measures general cognitive ability and the lower order factors of verbal, math, spatial, aircrew interest/aptitude, and perceptual speed (Ref 5,8).

The AFOQT is used to qualify civilians and prior-enlisted USAF personnel for officer commissioning through the OTS and ROTC programs. It is also used to qualify applicants who pass other educational and physical requirements for aircrew training (pilot, combat system officer, air battle manager, and remotely piloted aircraft pilot). The AFOQT has been validated for aircrew training (Ref 9-15) and for several other officer jobs (Ref 16-18).

Several studies have demonstrated the predictive validity of cognitive ability for pilot training performance (Ref 13,19). For example, Olea and Ree (Ref 15) compared the validity of general cognitive ability and specific abilities (including pilot job knowledge) for predicting several pilot training criteria in samples ranging from 1,867 to 3,942. General cognitive ability, specific abilities, and pilot job knowledge were measured by the AFOQT. The outcome criteria included academic grades, work samples of flight maneuvers, and an overall performance composite. Multiple correlations were compared to estimate the predictive efficiency of general ability and specific abilities for each criterion. Notwithstanding the apparent differences among the criteria, general ability was the best predictor, while specific abilities contributed only a little more. The validity for general ability ranged from .21 to .43 across all criteria, with a mean of .31. The incremental validity for the specific abilities beyond general intelligence ranged from

.07 to .14, with a mean of .10. These validities had been corrected for range restriction and are path model loadings. Most of the incremental validity for specific abilities came from the measurement of pilot job knowledge rather than specific abilities such as verbal, math, or spatial functioning.

An example of the practical utility of the AFOQT was provided by Duke and Ree (Ref 20). They demonstrated that scores on the AFOQT were related to the costs of pilot training through the number of hours it took to learn to fly. In a sample of 1,082 USAF pilot trainees, it was found that higher AFOQT scores translated into fewer flying hours required in the training aircraft. Costs per hour for flying the basic training and advanced training aircraft were obtained. The cost to train higher scoring trainees was less than the cost of training lower scoring trainees. Additionally, fewer hours spent learning to fly extended the useful life of the training aircraft by reducing physical stress and damage to the airframe.

In summary, there has been extensive research in the USAF on the use of cognitive ability tests, primarily the AFOQT. Interested readers are referred to Carretta and Ree (Ref 5, 19), who provide a thorough review of pilot selection methods including procedures for validation, potential problems encountered, and solutions to those problems.

2.1.2 Clinical Cognitive Testing. While accession procedures focus on intelligence, so, too, does much of the medical and other clinical testing of pilot candidates. The USAF Medical Flight Screening program screens pilot candidates prior to Undergraduate Pilot Training. In addition to several ophthalmic and cardiac diagnostic procedures, a number of psychological tests are administered (Ref 21,22). The primary purpose of the cognitive tests is to archive the individual's scores for future use. The intent is to develop a registry against which future testing might be compared. As such, the psychological portion of the program includes traditional measures of intelligence as well as newer computerized cognitive tasks.

As the primary purpose of the psychological testing is clinical, little work has focused on training. Indeed, the clinical tests used in the program have never been compared to pilot training outcomes. Boyd, Patterson, and Thompson (Ref 23) did, however, look at some of the tests against aircraft type later flown. Interestingly, this comparison may be a proxy for flight training outcomes. Usually, those highest in class rank are offered fighter aircraft and those lower are offered airlift/tanker aircraft. There are several issues that cloud this "hot hands get fighters" variable, such as the number of fighter training slots available at the time, the desire of the students, and Guard/Reserve pilots flying what their squadrons fly. However, the majority of the variance is probably accounted for by class rank.

Boyd, Patterson, and Thompson (Ref 23) compared one of the Medical Flight Screening intelligence tests, the Multidimensional Aptitude Battery (MAB) (Ref 24), and one of the personality tests, the NEO (Ref 25), to final airframe assignment. The three airframe types included fighter, bomber, and airlift/tanker. Bomber pilots come from either fighter or airlift/tanker training and, as such, are a mixed group. Fighter and airlift/tanker groups are more cleanly tracked as a function of performance in initial (T-37 or T-6) training, the first of the two-part pilot training protocol. There were significant differences between the groups on the MAB, with fighter pilots having intelligence quotients (IQs) of 2 to 3 point above the airlift/tanker pilots.

While these differences appear small, the fact of the matter is that there are relatively small differences across all pilots. Consequently, a couple of points out of a standard deviation of 7 points is really quite large. The Boyd et al. paper presented mean difference analysis of variance type statistics, so it is difficult to get a sense of the magnitudes of the differences or the effect sizes. Using the means and standard deviations in their Table A, as well as their sample sizes, it is possible to convert the differences found to a correlation statistic (Ref 26). The difference in verbal IQ between those assigned to fighters and those assigned to airlift/tankers was equivalent to a correlation of .14. The performance IQ difference was .15. The full-scale IQ difference was .18. These are all quite high, given what is discussed next, and suggest that intelligence, initial pilot training outcome, and subsequent airframe assignment are related.

2.1.3 Methodological Issues in Pilot Training Outcome Research. Several issues present themselves when engaged in this type of research (Ref 1). Perhaps the two most important have to do with the limited variability of both the intelligence scores and the outcomes.

The first issue comes from the fact that very little intellectual variance is left after initial pilot candidate selection. It is the intent of research such as this to show how intelligence differences across pilot candidates lead to differential outcomes. The trouble is that once someone has been selected to attend the Air Force Academy or scored high on the AFOQT, only quite intelligent students are left in training. Therefore, research participants do not represent the full range of intelligence found in the population at large. There are no 80-IQ, high school dropouts in the samples. This lack of range has the tendency to restrict the magnitude of correlations and other findings.

Second, the outcome measures lack variance and are of relatively low base rate. Since selection is so rigorous, relatively few pilot candidates actually fail training. The failure rate in the USAF is in the 10%-15% range. Further, the term “failure” is a misnomer. There are several reasons for not passing pilot training. The most obvious is due to flying training deficiency. This outcome is the closest to true failure. Other reasons, though, include medical problems, self-initiated elimination (DOR), and “Manifestation of Apprehension” (fear of flying). The issue of lack of variance and low base rate is compounded when the various reasons for failure are broken down into low single-digit percentages. This situation further restricts the potential magnitude of correlations.

Having constraints with both the predictors and the outcomes will lead to relatively weak relationships between what is left. The Hunter and Burke (Ref 27) and Martinussen (Ref 28) meta-analyses show that the typical correlation in these situations is probably only in the .11 to .13 range for uncorrected correlations.

2.2 Purpose

The purpose of this study was to determine the extent to which clinical cognitive functioning tests predict pilot training outcome. It is of further interest how each of the three different tests might differentially predict training outcome. This work not only focused on the “passing” versus “failing” of pilot training but also on additional, more focused, variables. For those “passing,” class rank, academic grades, daily flight grades, and check ride grades were used. For those “failing,” the reason for “failure” was analyzed, looking at flying training deficiency versus DOR. It was hoped that the use of three different clinical tests and multiple outcome variables would help to illuminate any relationships.

3.0 THE MULTIDIMENSIONAL APTITUDE BATTERY-II

The MAB (Ref 29) is a broad-based test of intellectual ability. It was patterned after the Wechsler Adult Intelligence Scales, the most widely used individually administered test of intelligence. Gignac (Ref 30) showed that the Wechsler and the MAB correlate highly. While the Wechsler is individually administered, the MAB can be given to groups and requires less total testing time and little time to score.

There have been two versions. The MAB was developed in 1984 (Ref 29). It was used quite early with USAF pilots by Retzlaff and Gibertini (Ref 31). The MAB was reviewed and restandardized in 1998 to ensure that it continued to be an effective measure of general cognitive ability. The result was the MAB-II (Ref 24). Most recently within the USAF, it has been shown to be useful with special operators by Chappelle, McDonald, Thompson, McMillan, and Marley (Ref 32).

The first version of the MAB was computerized by the USAF for the Neuropsychiatrically Enhanced Flight Screening program, the forerunner of the psychological portion of the current screening program (Ref 33). Retzlaff, King, and Callister (Ref 34) compared the original paper-and-pencil version of the MAB to the USAF computerized version and did not find significant differences between the two tests. The screening program subsequently used the computerized version published by the test author when it became available.

The MAB has 3 summary scores and 10 subtests. The test yields a full-scale IQ score, a verbal IQ score, and a performance IQ score. Verbal components are tapped by the Information, Comprehension, Arithmetic, Similarities, and Vocabulary subtests. Performance measures include the Digit Symbol Coding, Picture Completion, Spatial, Picture Arrangement, and Object Assembly subtests. Scores on each of the subtests are scaled to a mean of 50 and a standard deviation (SD) of 10. Full-scale, verbal, and performance scores are each scaled to a mean of 100 and an SD of 15. Reliabilities for the summary IQ scores range from .94 to .98 (Ref 24). Previous research has demonstrated that the full-scale score measures general cognitive ability in several age groups (Ref 35-40).

Carretta, Retzlaff, Callister, and King (Ref 39) examined the extent to which the AFOQT and the MAB measure the same constructs. A joint factor analysis revealed that both batteries had a hierarchical structure. The higher order factor in the AFOQT has been identified previously as general cognitive ability. The correlation between the higher order factors from the two batteries was .981, demonstrating that the general factors from both tests measure the same construct. The MAB verbal factor showed its highest between-battery correlation with the AFOQT verbal factor (.893) and its lowest correlation with aviation (.450). The MAB performance factor had its highest between-battery correlation with spatial (.854) and its lowest correlation with aviation (.587).

Table 1 presents the descriptions of the MAB subtests and summary IQ scores. As can be seen from the descriptions, the MAB-II is a very traditional and classic test of intelligence.

Table 1. Descriptions of the MAB-II Summary Scores and Subtests

Test	Description
Summary Scores	
Full-Scale IQ	Sum of verbal and performance scores
Verbal IQ	Sum of all verbal subtests
Performance IQ	Sum of all performance subtests
Verbal Subtests	
Information	Degree to which an examinee has amassed a body of knowledge about many topics
Comprehension	Measures "social acculturation," "social intelligence," and the conventional principles associated with moral and ethical standards
Arithmetic	The reasoning and solution to numeric and arithmetic problems
Similarities	A measurement of likenesses and differences of objects and their properties
Vocabulary	Identification of the meaning of words
Performance Subtests	
Digit Symbol	Measures visual motor activity in substituting symbols for digits
Picture Completion	Identification of pictures of common objects
Spatial	Two-dimensional visualization of abstract objects
Picture Arrangement	Measures ability to arrange pictures in an order that creates a meaningful story
Object Assembly	Ability to visualize complete objects from disassembled parts

Note: from Jackson (Ref 24).

3.1 Participants

Participants were 12,924 pilot training students. All were college graduates or were near completion of college. Of those reporting demographic information, 91% were male. Participants had a mean age of 23 years, and 99% were 30 years of age and under. Eighty-four percent reported that they were white. All participants were tested at either the USAF School of Aerospace Medicine (USAFSAM) or the U.S. Air Force Academy (USAFA).

3.2 Procedure

The MAB-II was administered to the pilot training students prior to entry into Undergraduate Pilot Training. Descriptive data (means and SDs) were computed for all scale scores. Univariate and multivariate statistics are presented comparing the clinical cognitive functioning test scores to outcome variables.

3.3 Outcomes

Training outcome data were from the first flying phase of USAF Undergraduate Pilot Training, which involved training in either the T-37 or T-6. These outcomes do not include advanced training in the T-38 or T-1 aircraft.

Several training performance outcome criteria were used. All participants had a final training outcome of "Pass" or "Fail." However, students may fail training for several reasons. We focused with individual analyses on those who failed due to poor flying performance (Flying Training Deficiency) (FlyDef) or who self-eliminated from training (DOR). Too few participants failed for other reasons such as medical problems or "Manifestation of Apprehension" to analyze these individually.

Several additional training performance criteria were available for students who successfully completed T-37/T-6 training: class rank, academic grades, daily flight grades, and check flight grades. Consequently, the seven variables were failure for all reasons, FlyDef, DOR, class rank, academic grades, daily flight grades, and check flight grades. Each was analyzed with t-tests and/or correlations as well as through multiple correlation procedures.

3.4 Results

Tables 2 through 7 contain the results for the analyses using the MAB and the criterion measures. Table 2 displays the means and SDs of the MAB for those who passed primary pilot training and those who failed for all reasons. As can be seen, the IQs of those who pass and those who fail are both quite high at about 120. For all 3 summary scores and all 10 subscales, graduates had higher mean scores than those who failed. All are statistically significant on t-test with the exception of the Vocabulary subtest. Point-biserial correlations are provided here as an effect size metric. While having a very large sample size is always welcomed by researchers, very small differences will usually be "statistically significant" yet may offer little actual practical predictive power. Indeed, that is the case here. Mean score differences between the two groups typically were only 1 point for the subtests and 2 points for the summary scales. The point-biserial correlations reinforce this issue with low correlations for the subtests and somewhat larger correlations for the summary scales. While 2-point differences on the summary scales may seem very small, the fact that the standard deviations are about 7 suggested that the magnitudes are not inconsequential. A correlation of .083 for the full-scale IQ score with the pass/fail criterion was observed and is interesting. Subsequent analyses using multivariate procedures improve these numbers. A caveat here is that the training failures included medical losses, so the group distinctions in this analysis are not as clear as one might like.

Table 2. Means and Standard Deviations for the MAB-II Scales for Pass and Fail

Subtest	Pass (N=11,579)		Fail[a] (N=1,345)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Summary Scores						
Full-Scale IQ	120.77	6.35	119.03	6.76	-9.48[b]	.083[b]
Verbal IQ	119.11	6.47	117.89	6.77	-6.52[b]	.057[b]
Performance IQ	119.70	8.02	117.60	8.70	-9.01[b]	.079[b]
Verbal Subtests						
Information	66.46	6.06	65.78	6.30	-3.92[b]	.034[b]
Comprehension	59.50	4.02	58.86	4.29	-5.52[b]	.048[b]
Arithmetic	61.09	6.53	59.49	6.52	-8.51[b]	.075[b]
Similarities	60.16	4.54	59.69	4.77	-3.58[b]	.031[b]
Vocabulary	59.50	6.71	59.18	7.11	-1.62	.014
Performance Subtests						
Digit Symbol	65.71	6.57	64.20	7.27	-7.91[b]	.069[b]
Picture Completion	60.14	6.13	58.88	6.48	-7.07[b]	.062[b]
Spatial	60.55	6.47	59.28	6.88	-6.81[b]	.060[b]
Picture Arrangement	52.34	7.08	51.69	7.35	-3.17[b]	.028[b]
Object Assembly	60.89	5.41	60.13	5.73	-4.87[b]	.043[b]

[a]"Fail" includes all reasons.
[b]$p<.01$.

Looking at passing versus failing for flying training deficiency (FlyDef) alone, Table 3 provides the means, standard deviation, and univariate tests. Again, for most subtests there is only about a 1-point difference between groups, and for the summary scores 2 to 3 points is seen. All mean score differences were significant with the exception of Information and Vocabulary on *t*-test. The magnitudes of the correlations were very similar to the prior analysis involving all those eliminated from training. Further, what seems to be developing with these analyses is a lack of specificity of score. The effects here seem to be quite general and without much variability across subtest content.

Table 3. Means and Standard Deviations for the MAB-II Scales for Pass and Failure Due to Flying Training Deficiency

Subtest	Pass (N=11,579)		FlyDef (N=559)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Summary Scores						
Full-Scale IQ	120.77	6.35	118.59	7.01	-7.91[a]	.072[a]
Verbal IQ	119.11	6.47	117.79	7.21	-4.68[a]	.042[a]
Performance IQ	119.70	8.02	116.83	8.94	-8.23[a]	.074[a]
Verbal Subtests						
Information	66.46	6.06	65.87	6.60	-2.25	.020
Comprehension	59.50	4.02	58.67	4.65	-4.74[a]	.043[a]
Arithmetic	61.09	6.53	59.00	6.49	-7.38[a]	.067[a]
Similarities	60.16	4.54	59.50	4.86	-3.30[a]	.030[a]
Vocabulary	59.50	6.71	59.76	7.39	0.88	-.008
Performance Subtests						
Digit Symbol	65.71	6.57	63.16	7.47	-8.90[a]	.081[a]
Picture Completion	60.14	6.13	58.70	6.72	-5.39[a]	.049[a]
Spatial	60.55	6.47	58.86	7.32	-6.01[a]	.054[a]
Picture Arrangement	52.34	7.08	51.47	7.40	-2.85[a]	.026[a]
Object Assembly	60.89	5.41	59.77	5.73	-4.76[a]	.043[a]

[a]$p<.01$.

Table 4 summarizes comparisons between graduates and those who DOR. While one would think that flying deficiency would be related to cognitive functioning, it would seem that requesting to be eliminated from training is more of a motivational or personality issue. Indeed, there is far less relationship between cognitive functioning and "failing" for this reason. Only half of the differences were statistically significant, and the magnitudes of the correlations were much smaller than the prior analyses. Looking at the scores that are significantly different, there does not appear to be a cohesive clinical theory explaining this finding. As such, it is likely that this was also a generalized effect and not variable specific.

Table 4. Means and Standard Deviations for the MAB-II Scales for Pass and Failure Due to "Drop on Request"

Subtest	Pass (N=11,579)		DOR (N=500)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Summary Scores						
Full-Scale IQ	120.77	6.35	119.52	6.46	-4.30[a]	.039[a]
Verbal IQ	119.11	6.47	118.39	6.06	-2.43	.022
Performance IQ	119.70	8.02	118.01	8.63	-4.59[a]	.042[a]
Verbal Subtests						
Information	66.46	6.06	66.26	5.85	-0.73	.007
Comprehension	59.50	4.02	59.07	3.78	-2.38	.022
Arithmetic	61.09	6.53	60.00	6.24	-3.67[a]	.033[a]
Similarities	60.16	4.54	60.16	4.69	0.01	.000
Vocabulary	59.50	6.71	58.99	6.67	-1.67	.015
Performance Subtests						
Digit Symbol	65.71	6.57	64.93	7.15	-2.60[a]	.024[a]
Picture Completion	60.14	6.13	59.11	6.37	-3.69[a]	.034[a]
Spatial	60.55	6.47	59.37	6.73	-4.01[a]	.036[a]
Picture Arrangement	52.34	7.08	51.97	7.26	-1.15	.010
Object Assembly	60.89	5.41	60.11	5.83	-3.15[a]	.029[a]

[a]$p<.01$.

For the purposes of comprehensiveness, Table 5 provides the comparison between those eliminated for FlyDef versus those for DOR. The flying deficiency group had lower scores on all variables than the DOR group. However, with the exception of Digit Symbol, none of these were statistically significant. Please note also that while a number of the correlations are at the same magnitude as prior analyses, here the reduced numbers of participants make these correlations not significant.

Table 5. Means and Standard Deviations for the MAB-II Subtests for Failure Due to Flying Training Deficiency and "Drop on Request"

Subtest	FlyDef (N=559)		DOR (N=500)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Summary Scores						
Full-Scale IQ	118.59	7.01	119.52	6.46	2.25	-.069
Verbal IQ	117.79	7.21	118.39	6.06	1.47	-.045
Performance IQ	116.83	8.94	118.01	8.63	2.19	-.067
Verbal Subtests						
Information	65.87	6.60	66.26	5.85	1.01	-.031
Comprehension	58.67	4.65	59.07	3.78	1.51	-.046
Arithmetic	59.00	6.49	60.00	6.24	2.53	-.078
Similarities	59.50	4.86	60.16	4.89	2.22	-.068
Vocabulary	59.76	7.39	58.99	6.67	-1.77	.054
Performance Subtests						
Digit Symbol	63.16	7.47	64.93	7.15	3.92[a]	-.120[a]
Picture Completion	58.70	6.72	59.11	6.37	1.00	-.031
Spatial	58.86	7.32	59.37	6.73	1.17	-.036
Picture Arrangement	51.47	7.40	51.97	7.26	1.12	-.034
Object Assembly	59.77	5.73	60.11	5.83	0.95	-.029

[a]$p<.01$.

The univariate point-biserial correlations from the prior tables are combined in Table 6 for easy comparison. Additionally, an ordinary least squares multiple regression is added to show the total predictive power of all the subscales combined. All 10 of the MAB subscales were "entered" into the equation. While logistic regression would generally be the preferred method for binomial outcomes such as this, ordinary least squares is used to more easily compare across the differing variable types in subsequent analyses. Oddly, FlyDef versus DOR has the highest multiple correlation (R) at 0.18, yet had subtests with the fewest univariate significant differences. This circumstance is one of the problems of dealing with small predictive relationships. The 0.11 multiple correlations found for pass versus fail and pass versus flying deficiency are probably more robust and are quite consistent with prior studies. These numbers are probably best viewed as "small but important."

Table 6. MAB-II Scale Point-Biserial Correlations for Failure and Reason for Failure

Subtest	Pass/ Fail	Pass/ FlyDef	Pass/ DOR	FlyDef/ DOR
Summary Scores				
Full-Scale IQ	.083[a]	.072[a]	.039[a]	-.069
Verbal IQ	.057[a]	.042[a]	.022	-.045
Performance IQ	.079[a]	.074[a]	.042[a]	-.067
Verbal Subtests				
Information	.034[a]	.020	.007	-.031
Comprehension	.048[a]	.043[a]	.022	-.046
Arithmetic	.075[a]	.067[a]	.033[a]	-.078
Similarities	.031[a]	.030[a]	.000	-.068
Vocabulary	.014	-.008	.015	.054
Performance Subtests				
Digit Symbol	.069[a]	.081[a]	.024[a]	-.120[a]
Picture Completion	.062[a]	.049[a]	.034[a]	-.031
Spatial	.060[a]	.054[a]	.036[a]	-.036
Picture Arrangement	.028[a]	.026[a]	.010	-.034
Object Assembly	.043[a]	.043[a]	.029[a]	-.029
Multiple R^b (10 subscales)	.108[a]	.111[a]	.056[a]	.178[a]

[a] $p<.01$.
[b] Regression only includes the 10 subscales.

Turning to the quality of passing, Table 7 provides correlations between the MAB scores and a number of course outcomes. Class rank is largely a function of academic grades, daily grades, and check rides, so the reader is warned of the correlated nature of these outcomes. Class rank, however, is probably the best single outcome. A correlation of 0.16 between full-scale IQ and class rank is strong. The multiple correlation of 0.21 is quite good. Academic grades seem to be particularly well modeled here with a multiple correlation of 0.27.

Table 7. MAB-II Scale Correlations for Training Performance

Subtest	Class Rank	Academic Grades	Daily Grades	Check Ride Grades
Summary Scores				
Full-Scale IQ	.157[a]	.233[a]	.124[a]	.110[a]
Verbal IQ	.123[a]	.224[a]	.084[a]	.083[a]
Performance IQ	.138[a]	.164[a]	.120[a]	.099[a]
Verbal Subtests				
Information	.065[a]	.130[a]	.048[a]	.041[a]
Comprehension	.091[a]	.154[a]	.074[a]	.066[a]
Arithmetic	.168[a]	.215[a]	.131[a]	.126[a]
Similarities	.043[a]	.118[a]	.017	.025
Vocabulary	.053[a]	.160[a]	.022[a]	.024
Performance Subtests				
Digit Symbol	.131[a]	.123[a]	.112[a]	.106[a]
Picture Completion	.090[a]	.130[a]	.073[a]	.058[a]
Spatial	.107[a]	.092[a]	.109[a]	.092[a]
Picture Arrangement	.049[a]	.076[a]	.029[a]	.018
Object Assembly	.080[a]	.127[a]	.066[a]	.056[a]
Multiple *R* (10 subscales)	.206[a]	.266[a]	.178[a]	.162[a]

[a]$p<.01$.

4.0 THE MICROCOG

The MicroCog (Ref 41) is a computerized test of cognitive functioning. It assesses a range of cognitive behaviors such as reaction time and memory. It was primarily developed to assess clinical pathology in patients. While the MAB is best viewed as a classic IQ test, the MicroCog comes more from a clinical neuropsychological perspective (Ref 42).

The test has 18 subtests, which result in 52 scores. The tasks include Timers, Address, Clocks, Story 1 Immediate Recall, Math, Tic Tac 1, Analogies, Numbers Forward, Story 2 Immediate Recall, Wordlists 1 and 2, Numbers Reversed, Address Delayed Recall, Object Match, Story 1 Delayed Recall, Alphabet, Tic Tac 2, Story 2 Delayed Recall, and Timers 2. The subtests are combined into five "domains" that include Attention/Mental Control, Memory, Reasoning/Calculation, Spatial Processing, and Reaction Time. It is unclear from the manual how the subtests were assigned to domains. The assignment of subtests could have been based on theory and/or on factor analysis.

Several higher order summary scores are derived. The first two, Information Processing: Speed and Information Processing: Accuracy, reflect a potential two-factor structure of the subtests. The second two summary scores purport to represent more general cognitive ability, where General Cognitive: Functioning is a function of the two Information Processing summary scores and General Cognitive: Proficiency is a summation of the Proficiency scores of all the subtests. Descriptions of the MicroCog indices as well as the subtests making up each index are displayed in Table 8.

Table 8. Descriptions of the MicroCog Summary Scores and Subtests

Index	Description	Subtests
Summary Scores		
Information Processing: Speed	Measures the time it takes an individual to complete simple and complex mental tasks	
Information Processing: Accuracy	Measures the accuracy of performance with no regard given to speed	
General Cognitive: Functioning	A measure of global cognitive functioning including equal weights of speed and accuracy index performance	
General Cognitive: Proficiency	A measure of global cognitive functioning including speed and accuracy index performance, with more weight given to accuracy	
Subtests		
Attention/Mental Control	Concentration, span of attention, diligence, persistence, resistance to interference	Numbers Forward Numbers Reversed Wordlists Alphabet
Memory	Short-term memory (storing information for a brief period) and long-term memory (storing information for a longer time period, from minutes to years)	Stories Immediate Stories Delayed Address Delayed Stories Time
Reasoning/Calculation	Inductive reasoning, cognitive flexibility, concept formation, basic arithmetic	Analogies Object Match Math
Spatial Processing	Memory for novel spatial arrangements, visuo-perceptual ability	Tic Tac Clocks
Reaction Time	Length of psychomotor time between presented stimulus and response, readiness to respond, vigilance, attention	Timers

The Information Processing and General Cognitive summary scores generally correlate with the Wechsler IQ test in the .50s. The manual (Ref 41) provides other validities for the domain scores. Here, for example, the MicroCog Memory Index correlates with the Wechsler Memory Scales in the .30s and .40s.

Chappelle, Ree, Barto, Teachout, and Thompson (Ref 43) compared the MAB and MicroCog in a structural equation model. They concluded that both tests have a factor representing general intelligence. Of interest, the MicroCog only produced one factor during the modeling. This finding suggests that while there may be five “domains” and four additional higher order summary scores, there is less specificity to the scores than a clinician or researcher may desire.

4.1 Participants

Participants were 5,582 pilot training students. As with the MAB, all were college graduates or were near completion of college. Of those reporting demographic information, 91% were male. Participants had a mean age of 23 years, and 99% were 30 years of age and under. Eighty-four percent reported that they were white. All participants were tested at either USAFSAM or USAFA.

4.2 Procedure

The MicroCog was administered to the pilot training students prior to entry into Undergraduate Pilot Training. As with the MAB, comparisons were made between those passing and failing T-37/T-6 training as well as against class performance. Univariate and multivariate statistics are presented comparing test scores to training performance variables.

4.3 Results

Tables 9 through 14 contain the results for the analyses using the MicroCog and the criterion measures. Table 9 presents the means and SDs for the graduates and those who failed for any reason. For all four summary scores and all five subtests, the graduates scored higher than the eliminees. All mean score comparisons were found to be statistically significant on t-test. Point-biserial correlations presented to model effect size were for the most part modest. Usually summary scores are more reliable and therefore more valid, but here Spatial Processing was high at 0.11.

Only those failing for FlyDef reasons are included in Table 10. The findings parallel those of the analyses for all graduates and eliminees. Again, modest differences were found.

Looking only at those requesting to be eliminated (DOR) in Table 11, few differences were seen. Again, it is probable that DOR is driven by variables in addition to intellectual ability. Interestingly, as before, Spatial Processing was the most predictive. One wonders if, perhaps, there is some combination of motivation, personality, and very specific cognitive ability that results in the request to be eliminated.

As with the MAB, the DOR group had higher scores across the MicroCog scores than the FlyDef group (Table 12). Also, as with the MAB, the reduced numbers of participants resulted in only about half of these differences being statistically significant. The results still suggest some sort of generalized intelligence effect as opposed to a cognitive function specific effect.

Table 9. Means and Standard Deviations for the MicroCog Scales by Pass and Fail

Subtest	Pass (N=4,992)		Fail[a] (N=590)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Summary Scores						
Information Processing: Speed	105.57	12.01	102.89	12.49	-5.09[b]	.068[b]
Information Processing: Accuracy	98.32	12.62	96.15	13.34	-3.93[b]	.053[b]
General Cognitive: Functioning	106.75	14.28	102.89	13.94	-6.22[b]	.083[b]
General Cognitive: Proficiency	104.22	9.97	101.26	9.90	-6.84[b]	.091[b]
Subtests						
Attention/Mental Control	103.34	12.13	100.77	12.79	-4.85[b]	.065[b]
Memory	109.22	13.67	107.50	14.30	-2.87[b]	.038[b]
Reasoning/Calculation	97.56	12.64	95.93	12.98	-2.95[b]	.039[b]
Spatial Processing	107.25	9.45	103.77	10.93	-8.32[b]	.111[b]
Reaction Time	99.52	12.00	97.16	13.05	-4.47[b]	.060[b]

[a]"Fail" includes all reasons.
[b]$p<.01$.

Table 10. Means and Standard Deviations for the MicroCog Scales for Pass and Failure Due to Flying Training Deficiency

Subtest	Pass (N=4,992)		FlyDef (N=246)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Summary Scores						
Information Processing: Speed	105.57	12.01	101.15	12.72	-5.62[a]	.077[a]
Information Processing: Accuracy	98.32	12.62	94.55	12.96	-4.57[a]	.063[a]
General Cognitive: Functioning	106.75	14.28	100.19	13.44	-7.06[a]	.097[a]
General Cognitive: Proficiency	104.22	9.97	99.39	9.73	-7.43[a]	.102[a]
Subtests						
Attention/Mental Control	103.34	12.13	97.98	12.76	-6.76[a]	.093[a]
Memory	109.22	13.67	105.42	14.08	-4.24[a]	.059[a]
Reasoning/Calculation	97.56	12.64	94.77	14.16	-3.36[a]	.046[a]
Spatial Processing	107.25	9.45	102.67	10.60	-7.38[a]	.101[a]
Reaction Time	99.52	12.00	95.00	13.73	-5.74[a]	.079[a]

[a]$p<.01$.

Table 11. Means and Standard Deviations for the MicroCog Scales for Pass and Failure Due to "Drop on Request"

Subtest	Pass (N=4,992)		DOR (N=202)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Summary Scores						
Information Processing: Speed	105.57	12.01	104.06	12.91	-1.74	.024
Information Processing: Accuracy	98.32	12.62	97.32	13.88	-1.10	.015
General Cognitive: Functioning	106.75	14.28	104.84	14.34	-1.86	.026
General Cognitive: Proficiency	104.22	9.97	102.65	9.98	-2.20	.031
Subtests						
Attention/Mental Control	103.34	12.13	101.97	12.41	-1.57	.022
Memory	109.22	13.67	109.00	13.93	-0.22	.003
Reasoning/Calculation	97.56	12.64	97.21	12.30	-0.38	.005
Spatial Processing	107.25	9.45	104.84	11.11	-3.53[a]	.049[a]
Reaction Time	99.52	12.00	98.57	12.49	-1.10	.015

[a]p<.01.

Table 12. Means and Standard Deviations for the MicroCog Scales by Failure Due to Flying Training Deficiency and "Drop on Request"

Subtest	FlyDef (N=246)		DOR (N=202)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Summary Scores						
Information Processing: Speed	101.15	12.72	104.06	12.91	2.40	-.133
Information Processing: Accuracy	94.55	12.96	97.32	13.88	2.18	-.103
General Cognitive: Functioning	100.19	13.44	104.84	14.34	3.54[a]	-.165[a]
General Cognitive: Proficiency	99.39	9.73	102.65	9.98	3.48[a]	-.163[a]
Subtests						
Attention/Mental Control	97.98	12.76	101.97	12.41	3.34[a]	-.156[a]
Memory	105.42	14.08	109.00	13.93	2.69[a]	-.126[a]
Reasoning/Calculation	94.77	14.16	97.21	12.30	1.93	-.091
Spatial Processing	102.67	10.60	104.84	11.11	2.11	-.099
Reaction Time	95.00	13.73	98.57	12.49	2.86[a]	-.134[a]

[a]p<.01.

Table 13 provides the point-biserial correlations from the prior tables in a comparison format. Additionally, multiple regressions are presented to model the maximal predictive power of the MicroCog. Since the summary scores are not clearly hierarchical of the subtests, all summary scores and subtests were included in the full regression model. The $R = 0.13$ and 0.14 multiple correlations for pass versus fail and pass versus flying deficiency were certainly consistent with the magnitude of results observed in prior meta-analyses. While an R of 0.20 was seen for flying FlyDef versus DOR, the small sample size led this to be nonsignificant.

Correlations between the MicroCog scores and class performance for those passing are in Table 14. The 0.23 and 0.25 multiple correlations against class rank and academic grades are actually quite impressive given the constrained variance. The 0.19 and 0.17 for daily flight grades and check rides grades are nicely supportive.

Table 13. MicroCog Scale Point-Biserial Correlations with Failure and Reason for Failure

Subtest	Pass/ Fail	Pass/ FlyDef	Pass/ DOR	FlyDef/ DOR
Summary Scores				
Information Processing: Speed	.068[a]	.077[a]	.024	-.133
Information Processing: Accuracy	.053[a]	.063[a]	.015	-.103
General Cognitive: Functioning	.083[a]	.097[a]	.026	-.165[a]
General Cognitive: Proficiency	.091[a]	.102[a]	.031	-.163[a]
Subtests				
Attention/Mental Control	.065[a]	.093[a]	.022	-.156[a]
Memory	.038[a]	.059[a]	.003	-.126[a]
Reasoning/Calculation	.039[a]	.046[a]	.005	-.091
Spatial Processing	.111[a]	.101[a]	.049[a]	-.099
Reaction Time	.060[a]	.079[a]	.015	-.134[a]
Multiple R[b]	.126[a]	.138[a]	.057	.195

[a]p<.01.
[b]All four summary scores and all five subtests were entered into the multiple regression.

Table 14. MicroCog Scale Correlations with Training Performance

Subtest	Class Rank	Academic Grades	Daily Grades	Check Rides
Summary Scores				
Information Processing: Speed	.139[a]	.075[a]	.137[a]	.102[a]
Information Processing: Accuracy	.138[a]	.220[a]	.093[a]	.097[a]
General Cognitive: Functioning	.202[a]	.206[a]	.165[a]	.149[a]
General Cognitive: Proficiency	.201[a]	.204[a]	.170[a]	.148[a]
Subtests				
Attention/Mental Control	.138[a]	.146[a]	.107[a]	.098[a]
Memory	.174[a]	.169[a]	.138[a]	.134[a]
Reasoning/Calculation	.120[a]	.147[a]	.100[a]	.075[a]
Spatial Processing	.146[a]	.114[a]	.134[a]	.116[a]
Reaction Time	.081[a]	.074[a]	.070[a]	.058[a]
Multiple R^b	.231[a]	.249[a]	.192[a]	.173[a]

[a] $p<.01$.
[b] All four summary scores and all five subtests were entered into the multiple regression.

5.0 THE COGSCREEN

The CogScreen (Ref 44) is a test of cognitive ability intended for use in the assessment of pilots. While the MAB is a test of relatively complex, higher order intellectual processes, the CogScreen tasks generally involve more fundamental processes such as reaction time. Its developers claim that it taps abilities necessary in the performance of aviation duties and was supported by the Federal Aviation Administration as a measure of underlying abilities related to flying. The development and normative sample consists of 584 commercial aviators.

The CogScreen has a number of tasks that result in 65 scores. The tasks include Math, Visual Sequence Comparison, Matching-to-Sample, Manikin, Divided Attention, Auditory Sequence Comparison, Pathfinder, Shifting Attention, and Dual Task. Table 15 provides descriptions of the CogScreen subtests. Each of the tasks is scored in several ways. Typical scorings include task speed, accuracy, and throughput. Throughput is a function of speed and accuracy, reflecting the number of correct responses per minute. It is indicative of the amount of work accomplished. Several tasks also include process completion measures, which quantify task specific behavior such as control of the computer screen elements.

Table 15. Description of the CogScreen Subtests[a]

Subtest	Definition
Math	Calculate multistep word problems.
Visual Sequence Comparison	Determine whether two alphanumeric strings presented side-by-side are the same or different.
Matching-to-Sample	After viewing a four-by-four grid pattern, select the correct pattern from two grids displayed side by side.
Manikin	Determine in which hand a figure is holding a flag by mentally rotating the image in one of four positions.
Divided Attention	Monitor the vertical movement of a cursor within a circle and return it to center when it exceeds the boundaries. The task is performed alone and with the Visual Sequence Comparison task.
Auditory Sequence Comparison	Compare two series of four to eight tones of varying pitch presented sequentially.
Pathfinder	Determine which character comes next in a series after being presented with three sequencing rules of the characters (numbers, letters, or both).
Shifting Attention	Determine the sequence of letters and numbers based upon changing rules.
Dual Task	Perform a tracking test and a delayed recall memory task separately, then at the same time.

[a]From Kay (Ref 44).

Stability of the CogScreen was on 199 airline pilots retested at 6 and 12 months after initial test administration (Ref 44). Throughput variables were selected for reliability estimation because they have normal distributions and are a combination of speed and accuracy measures. Test-retest reliability coefficients for throughput measures ranged from .69 to .90, with an average coefficient of .80. For the speed scores, reliability coefficients ranged from .63 to .91, with an average coefficient of .80. Reliability estimates were not calculated for accuracy and process variables because of the low variability in scores (Ref 44).

5.1 Participants

Participants were 7,003 pilot training students. As with the MAB and MicroCog, all were college graduates or were near completion of college. Of those reporting demographic information, 91% were male. Participants had a mean age of 23 years, and 99% were 30 years of age and under. Eighty-four percent reported that they were white. All participants were tested at either USAFSAM or USAFA.

5.2 Procedure

The CogScreen was administered to the pilot training students prior to entry into Undergraduate Pilot Training. Because of the large number of scores derived from the CogScreen, only the throughput scores were examined here. Those variables should represent the vast majority of the CogScreen's assessment ability. Limiting the number of variables also improves statistical robustness by increasing the ratio of subjects to variables. Univariate and multivariate statistics are presented comparing test scores to training performance.

5.3 Results

Tables 16 through 21 contain the results for the analyses using the Cogscreen and the criterion measures. Table 16 displays the descriptive statistics for pilot training graduates and and those not graduating. Further, both t-tests and point-biserial correlations are provided for significance testing and effect size estimation. Only about half of the mean differences were statistically significant, and the magnitude of the correlations was modest. The biggest difference was observed for one of the Shifting Attention subtests. Neither the MAB nor the MicroCog have such a subtest. Military pilot training would certainly seem to involve the shifting of attention.

Unlike analyses with the MAB and MicroCog, separating out only the failures due to flying seems to improve the predictive power of the CogScreen. As shown in Table 17, an additional number of subscales became significant. The magnitude of the correlations also improved slightly.

Table 18 provides the analysis for those seeking to self-eliminate. Only one of the many variables is seen as having statistically significant differences. As with the prior tests, DOR seems not to be driven by intellectual capability.

Only a few statistically significant differences are seen in Table 19 between those eliminated for flying deficiency reasons and those self-eliminating. Where there are differences, DOR participants have higher scores than those leaving for flying reasons.

Table 16. Means and Standard Deviations for the CogScreen Subtests for Pass and Fail

Subtest	Pass (N=6,265)		Fail[a] (N=738)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Math	2.30	1.04	2.17	.91	-3.29[b]	.039[b]
Visual Sequence Comparison	30.73	6.72	30.02	6.75	-2.70[b]	.032[b]
Match-to-Sample	50.36	10.60	49.56	10.35	-1.95	.023
Manikin	35.39	8.94	34.33	9.04	-3.05[b]	.036[b]
Divided Attention	28.79	7.32	28.33	7.48	-1.60	.019
Auditory Sequence	90.36	25.23	87.01	23.77	-3.42[b]	.041[b]
Pathfinder: Num[b]er	83.20	18.78	81.92	19.07	-1.74	.021
Pathfinder: Letter	88.26	18.33	88.16	19.82	-0.13	.002
Pathfinder: Combination	61.84	15.88	60.71	16.79	-1.82	.022
Shifting Attention: Direction	110.28	21.16	110.09	20.73	-0.23	.003
Shifting Attention: Color	100.32	17.54	99.46	17.80	-1.26	.015
Shifting Attention: Instruction	87.47	29.20	84.45	17.45	-2.75[b]	.033[b]
Shifting Attention: Discovery	54.70	17.80	51.49	15.99	-4.69[b]	.056[b]
Dual Task: Alone	170.37	206.17	157.93	214.18	-1.54	.018
Dual Task: Dual	132.46	191.54	116.43	70.42	-2.26[b]	.027

[a]"Fail" includes all reasons.
[b]$p<.01$.

Table 17. Means and Standard Deviations for the CogScreen Subtests for Pass and Failure Due to Flying Training Deficiency

Subtest	Pass (N=6,265)		FlyDef (N=305)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Math	2.30	1.05	2.08	0.89	-3.67[a]	.045[a]
Visual Sequence Comparison	30.73	6.72	29.30	6.55	-3.64[a]	.045[a]
Match-to-Sample	50.36	10.60	48.42	9.86	-3.13[a]	.039[a]
Manikin	35.39	8.94	33.13	8.97	-4.31[a]	.053[a]
Divided Attention	28.79	7.32	27.66	7.14	-2.61[a]	.032[a]
Auditory Sequence	90.36	25.23	85.35	21.84	-3.40[a]	.042[a]
Pathfinder: Number	83.20	18.78	80.36	17.63	-2.58	.032[a]
Pathfinder: Letter	88.26	18.33	87.89	18.00	-0.34	.004
Pathfinder: Combination	61.84	15.88	59.85	15.76	-2.15	.026
Shifting Attention: Direction	110.28	21.16	107.96	19.20	-1.88	.023
Shifting Attention: Color	100.32	17.54	97.86	16.43	-2.40	.030
Shifting Attention: Instruction	87.47	29.20	81.96	17.08	-3.27[a]	.040[a]
Shifting Attention: Discovery	54.70	17.80	50.41	15.90	-4.13[a]	.051[a]
Dual Task: Alone	170.37	206.17	165.50	327.01	-0.39	.005
Dual Task: Dual	132.46	191.54	112.19	75.98	-1.84	.023

[a]$p<.01$.

Table 18. Means and Standard Deviations for the CogScreen Subtests for Pass and Failure Due to "Drop on Request"

Subtest	Pass (N=6,265)		DOR (N=291)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Math	2.30	1.05	2.20	0.92	-1.63	.020
Visual Sequence Comparison	30.73	6.72	30.44	6.93	-0.73	.009
Match-to-Sample	50.36	10.60	50.78	10.93	0.65	-.088
Manikin	35.39	8.94	35.35	9.13	-0.07	.001
Divided Attention	28.79	7.32	28.83	7.64	0.10	-.001
Auditory Sequence	90.36	25.23	88.96	24.43	-0.93	.011
Pathfinder: Number	83.20	18.78	82.70	20.24	-0.44	.005
Pathfinder: Letter	88.26	18.33	87.09	19.88	-1.06	.013
Pathfinder: Combination	61.84	15.88	61.40	17.24	-0.47	.006
Shifting Attention: Direction	110.28	21.16	111.59	22.32	1.03	-.013
Shifting Attention: Color	100.32	17.54	100.93	19.89	0.57	-.007
Shifting Attention: Instruction	87.47	29.20	86.63	18.80	-0.49	.006
Shifting Attention: Discovery	54.70	17.80	51.60	16.36	-2.91[a]	.036[a]
Dual Task: Alone	170.37	206.17	152.73	58.16	-1.46	.018
Dual Task: Dual	132.46	191.54	121.35	73.11	-0.99	.012

[a]p<.01.

Table 19. Means and Standard Deviations for the CogScreen Subtests for Failure Due to Flying Training Deficiency and "Drop On Request"

Subtest	FlyDef (N=305)		DOR (N=291)		Univariate Analysis	
	Mean	SD	Mean	SD	*t*-test	*r*
Math	2.08	0.89	2.20	0.92	1.65	-.067
Visual Sequence Comparison	29.30	6.55	30.44	6.93	2.06	-.084[a]
Match-to-Sample	48.42	9.86	50.78	10.93	2.76[a]	-.113[a]
Manikin	33.13	8.97	35.35	9.13	2.99[a]	-.122[a]
Divided Attention	27.66	7.14	28.83	7.64	1.92	-.079
Auditory Sequence	85.35	21.84	88.96	24.43	1.90	-.078
Pathfinder: Number	80.36	17.63	82.70	20.24	1.51	-.062
Pathfinder: Letter	87.89	18.00	87.09	19.88	-0.52	.021
Pathfinder: Combination	59.85	15.76	61.40	17.24	1.15	-.047
Shifting Attention: Direction	107.96	19.20	111.59	22.32	2.13	-.087
Shifting Attention: Color	97.86	16.43	100.93	19.89	2.06	-.084
Shifting Attention: Instruction	81.96	17.08	86.63	18.80	3.18[a]	-.129[a]
Shifting Attention: Discovery	50.41	15.90	51.60	16.36	0.90	-.037
Dual Task: Alone	165.50	327.01	152.73	58.16	-0.66	.027
Dual Task: Dual	112.19	75.98	121.35	73.11	1.50	-.061

[a]p<.01.

Combining the point-biserial correlations from the prior tables with multiple regressions in Table 20, comparisons across variables and outcomes are possible. Both the univariate and the multiple correlations were modest.

Training performance results for those who passed training are presented in Table 21. In general, the relationships here are stronger than the relationships with reasons for failure. The multiple correlation of 0.17 with class rank was consistent with prior results. The Math and Shifting Attention tasks appear to be the most predictive.

Table 20. CogScreen Subtest Point-Biserial Correlations with Failure and Reason for Failure

Subtest	Pass/ Fail	Pass/ FlyDef	Pass/ DOR	FlyDef/ DOR
Math	.039[a]	.045[a]	.020	-.067
Visual Sequence Comparison	.032[a]	.045[a]	.009	-.084[a]
Match-to-Sample	.023	.039[a]	-.088	-.113[a]
Manikin	.036[a]	.053[a]	.001	-.122[a]
Divided Attention	.019	.032[a]	-.001	-.079
Auditory Sequence	.041[a]	.042[a]	.011	-.078
Pathfinder: Number	.021	.032[a]	.005	-.062
Pathfinder: Letter	.002	.004	.013	.021
Pathfinder: Combination	.022	.026	.006	-.047
Shifting Attention: Direction	.003	.023	-.013	-.087
Shifting Attention: Color	.015	.030	-.007	-.084
Shifting Attention: Instruction	.033[a]	.040[a]	.006	-.129[a]
Shifting Attention: Discovery	.056[a]	.051[a]	.036[a]	-.037
Dual Task: Alone	.018	.005	.018	.027
Dual Task: Dual	.027	.023	.012	-.061
Multiple R^b	.089[a]	.093[a]	.061	.243[a]

[a] $p<.01$.
[b] All subtests were entered into the multiple regression.

Table 21. CogScreen Subtest Correlations with Training Performance

Subtest	Class Rank	Academic Grades	Daily Grades	Check Rides
Math	.125[a]	.139[a]	.079[a]	.092[a]
Visual Sequence Comparison	.052[a]	.026	.029	.027
Match-to-Sample	.060[a]	.011	.054[a]	.042[a]
Manikin	.098[a]	.017	.082[a]	.083[a]
Divided Attention	.036[a]	.030	.017	.017
Auditory Sequence	.052[a]	.014	.037[a]	.025
Pathfinder: Number	.057[a]	.049[a]	.028	.023
Pathfinder: Letter	.045[a]	.029	.012	.011
Pathfinder: Combination	.071[a]	.086[a]	.029	.036
Shifting Attention: Direction	.051[a]	.018	.030	.031
Shifting Attention: Color	.051[a]	.037[a]	.029	.026
Shifting Attention: Instruction	.069[a]	.044[a]	.051[a]	.041[a]
Shifting Attention: Discovery	.099[a]	.095[a]	.075[a]	.077[a]
Dual Task: Alone	.007	-.101	-.006	.007
Dual Task: Dual	.018	.006	.019	.001
Multiple R^b	.170[a]	.173[a]	.134[a]	.135[a]

[a]p<.01.
[b]All subtests were entered into the multiple regression.

6.0 DISCUSSION

This study examined the relationship between clinical tests of cognitive functioning and the results of USAF primary pilot training. The tests had been administered for medical purposes prior to the beginning of training. Three tests were analyzed including the Multidimensional Aptitude Battery, the MicroCog, and the CogScreen. Several outcomes were used to model pilot training performance, including the overall variable of passing versus failing training. For students failing training, the reason for elimination was examined including flying training deficiency and being “Dropped on Request.” For students passing training, four variables were examined: class rank, academic grades, daily grades, and check ride grades.

Overall, the results were consistent with prior work showing that the limited variance among these students would result in uncorrected correlations in the low teens.

The three tests showed similar predictiveness. The MicroCog probably showed the best ability to predict outcome, followed by the MAB, followed by the CogScreen. As was suggested by Olea and Ree (Ref 15), little subscale specificity was found. Broad and general intellectual functioning was seen to be at work in this study. While the three clinical tests had several subscales and differed in focus, no subscale or specific intellectual function stood out as more predictive than another.

With regard to the outcome variables, the prediction of how well someone does who passes pilot training appears to be more predictable than who will fail pilot training. Indeed, some multiple correlations in the mid-twenties were found in the prediction of class rank and academic grades.

Failing pilot training is a very heterogeneous experience. Our initial analyses included students eliminated for all reasons, including medical and Manifestation of Anxiety. Focusing specifically on flying training deficiency and "Drop on Request" students showed small but important relationships. In general, cognitive ability appeared to be related more closely to elimination due to flying training deficiency than to self-initiated elimination.

While only small failure predictions were found and modest class performance predictions were seen, these two outcome classes are combinable. The tests predicted both how well someone succeeds and who will fail. As such, the tests are probably predicting more than is modeled by either class of analysis alone. A multiple correlation for failure of 0.15 and a multiple correlation for class rank of 0.20 add to something greater.

There are always limitations to studies of pilot training outcome. This study is no different. It is surprising how many participants are needed to model all of the various reasons for being eliminated from pilot training. Although there were thousands of participants in the current study, there were still too few to have sufficient numbers of participants to analyze the lower base rate reasons for elimination such as medical removal and Manifestations of Apprehension.

We see two lines of work going forward. The first would be to look at advanced training assignment similar to the work of Boyd, Patterson, and Thompson (Ref 23) and to look at advanced training performance in the T-38 and T-1 tracks. Here the numbers of failures are so low that only the class performance variables of class rank, academic grades, daily flying grades, and check rides are probably relevant. These outcomes would add to our knowledge about the overall validities of the tests: a little failure prediction added to some primary class performance prediction and, finally, added to some advanced class performance prediction.

The second line of work that would be interesting would be to add AFOQT and Pilot Candidate Selection Method (PCSM) scores to this dataset. The AFOQT measures cognitive ability (like the MAB, MicroCog, and CogScreen), but also includes aviation knowledge. The PCSM includes other measures shown to be related to pilot training performance (psychomotor, flying experience) not measured by the clinical tests. It would be of interest to determine whether any of the clinical tests adds to the predictiveness of the AFOQT and PCSM. If they do not demonstrate incremental validity, then it is quite certain that we are truly dealing with a very generalized intellectual process.

Finally, from a methodological perspective, the current study has taken a very conservative approach to the analyses of these data. It is common, depending upon viewpoint, to "correct" the data for various reasons. Specifically, the data could be corrected for range restriction due to prior selection of the students and unreliability of the training criteria. For analyses involving the pass/fail training scores, the correlations could also be corrected for dichotomization of the criteria.

7.0 REFERENCES

1. Paullin C, Katz L, Bruskiewicz KT, Houston J, Damos D, *Review of Aviator Selection*, Technical Report 1183, U.S. Army Research Institute for the Behavioral and Social Sciences, Arlington, VA, Jul 2006.

2. North RA, Griffin GR, *Aviator Selection 1919-1977*, NAMRL SR 77-2, Naval Aerospace Medical Research Laboratory, Naval Air Station, Pensacola, FL, Oct 1977.

3. Carretta TR, Ree MJ, "Factor Structure of the Air Force Officer Qualifying Test: Analysis and Comparison," *Military Psychology*, **8**(1), 1996, pp. 29-42.

4. Howse WR, Damos DL, *A Bibliographic Database for the History of Pilot Training Selection*, AFCAPS-FR-2011-0010, Air Force Personnel Center, Randolph AFB, TX, Jul 2011.

5. Carretta TR, Ree MJ, "General and Specific Cognitive and Psychomotor Abilities in Personnel Selection: The Prediction of Training and Job Performance," *International Journal of Selection and Assessment*, **8**, 2000, pp. 227-36.

6. Carretta TR, *Development and Validation of the Test of Basic Aviation Skills (TBAS),* AFRL-HE-WP-TR-2005-0172, Air Force Research Laboratory, Human Effectiveness Directorate, Wright-Patterson AFB, OH, Nov 2005.

7. Skinner J, Ree MJ, *Air Force Officer Qualifying Test (AFOQT): Item and Factor Analysis of Form O*, AFHRL-TR-86-68, Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB, TX, Aug 1987.

8. Drasgow F, Nye CD, Carretta TR, Ree MJ, "Factor Structure of the Air Force Officer Qualification Test Form S: Analysis and Comparison with Previous Forms," *Military Psychology*, **22**, 2010, pp. 68−85.

9. Arth TO, Steuck KW, Sorrentino CT, Burke EF, *Air Force Officer Qualifying Test (AFOQT): Predictors of Undergraduate Pilot Training and Undergraduate Navigator Training Success*, AFHRL-TP-89-52, Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB, TX, May 1990.

10. Carretta TR, "Recent Developments in U. S. Air Force Pilot Candidate Selection and Classification," *Aviation, Space, and Environmental Medicine*, **63**, 1992, pp. 1112-4.

11. Carretta TR, Ree MJ, "Air Force Officer Qualifying Test Validity for Predicting Pilot Training Performance," *Journal of Business and Psychology*, **9**, 1995, pp. 379-88.

12. Carretta TR, *Predictive Validity of the Air Force Officer Qualifying Test for USAF Air Battle Manager Training Performance*, AFRL-RH-WP-TR-2009-0007, Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Interface Division, Wright-Patterson AFB, OH, Sep 2008.

13. Ree MJ, Carretta TR, "Central Role of g in Military Pilot Selection," *International Journal of Aviation Psychology*, **6**, 1996, pp. 111-23.

14. Ree MJ, Carretta TR, Teachout MS, "Role of Ability and Prior Knowledge in Complex Training Performance," *Journal of Applied Psychology*, **80**, 1995, pp. 721-30.

15. Olea MM, Ree MJ, “Predicting Pilot and Navigator Criteria: Not Much More than g,” *Journal of Applied Psychology*, **79**, 1994, pp. 845-51.

16. Carretta TR, “Predictive Validity of the Air Force Officer Qualifying Test for Non-Rated Officer Specialties,” *Military Psychology*, **22**, 2010, pp. 450-64.

17. Arth TO, Skinner MJ, “Aptitude Selectors for Air Force Officer Non-Aircrew Jobs,” in *Proceedings of the 28th Annual Conference of the Military Testing Association*, New London, CT, 3-7 Nov 1986, pp. 301-6.

18. Finegold LS, Rogers D, *Relationship Between Air Force Officer Qualifying Test Scores and Success in Air Weapons Controller Training*, AFHRL-TR-85-13, Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB, TX, Jun 1985.

19. Carretta TR, Ree MJ, “Pilot Selection Methods,” in Tsang PS, Vidulich MA, eds., **Human factors in Transportation: Principles and Practice of Aviation Psychology**, Lawrence Erlbaum Associates, Inc., Publishers, Mahwah, NJ, 2003, pp. 357-96.

20. Duke AP, Ree MJ, “Better Candidates Fly Fewer Training Hours: Another Time Testing Pays Off,” *International Journal of Selection and Assessment*, **4**, 1996, pp. 115-21.

21. King RE, Barto E, Ree MJ, Teachout MS, *Compilation of Pilot Personality Norms*, AFRL-SA-WP-TR-2011-0008, U.S. Air Force School of Aerospace Medicine, Wright-Patterson AFB, OH, Jul 2011.

22. King RE, Barto E, Ree MJ, Teachout MS, Retzlaff PD, *Compilation of Pilot Cognitive Ability Norms,* AFRL-SA-WP-TR-2012-0001, U.S. Air Force School of Aerospace Medicine, Aeromedical Research Department, Wright-Patterson Air Force Base, OH, Dec 2011.

23. Boyd JE, Patterson JC, Thompson BT, “Psychological Test Profiles of USAF Pilots Before Training vs. Type Aircraft Flown,” *Aviation, Space, and Environmental Medicine*, **76**, 2005, pp. 463-8.

24. Jackson, DN, **Multidimensional Aptitude Battery-II: Manual**. SIGMA Assessment Systems, Port Huron, MI, 1998.

25. Costa PT, McCrae RR, **The NEO Personality Inventory Manual**, Psychological Assessment Resources, Odessa, FL, 1985.

26. Lipsey MW, Wilson DB, **Practical Meta-Analysis**, Sage, Thousand Oaks, CA, 2000.

27. Hunter DR, Burke EF, “Predicting Aircraft Pilot-Training Success: A Meta-Analysis of Published Research,” *International Journal of Aviation Psychology*, **4**, 1994, pp. 297-313.

28. Martinussen, M. “Psychological Measures as Predictors of Pilot Performance: A Meta-Analysis,” *International Journal of Aviation Psychology*, **6**, 1996, pp. 1-20.

29. Jackson DN, **Multidimensional Aptitude Battery: Manual**, Research Psychologists Press, Port Huron, MI, 1984.

30. Gignac GE, “A Confirmatory Examination of the Factor Structure of the Multidimensional Aptitude Battery: Contrasting Oblique, Higher Order, and Nested Factor Models,” *Educational and Psychological Measurement,* **66**, 2006, pp. 136-45.

31. Retzlaff PD, Gibertini M, “Objective Psychological Testing of U.S. Air Force Officers in Pilot Training,” *Aviation, Space, and Environmental Medicine*, **59**(7), Jul 1988, pp. 661-3.

32. Chappelle W, McDonald K, Thompson W, McMillan K, Marley M, *Multiple Aptitude Battery-II Normative Intelligence Test Data that Distinguish U.S. Air Force AC-130 Gunship Sensor Operators,* AFRL-SA-BR-TR-2010-0006, U.S. Air Force School of Aerospace Medicine, Brooks City-Base, TX, Jun 2010.

33. King RE, Flynn CF, “Defining and Measuring the ‘Right Stuff’: Neuropsychiatrically Enhanced Flight Screening (N-EFS),” *Aviation, Space, and Environmental Medicine*, **66**(1), Oct 1995, pp. 951-6.

34. Retzlaff PD, King RE, Callister JD, *USAF Pilot Training Completion and Retention: A Ten Year Follow-Up on Psychological Testing*, Technical Report AL/AO-TR-1995-0124, Armstrong Laboratory, Brooks AFB, TX, Aug 1995.

35. Wallbrown FH, Carmin CN, Barnett RW, “Investigating the Construct Validity of the Multidimensional Aptitude Battery,” *Psychological Reports*, **62**(3), Jun 1988, pp. 871-8.

36. Wallbrown FH, Carmin CN, Barnett RW, “A Further Note on the Construct Validity of the Multidimensional Aptitude Battery,” *Journal of Clinical Psychology*, **45**(3), May 1989, pp. 429-33.

37. Lee MS, Wallbrown FH, Blaha J, “Note on the Construct Validity of the Multidimensional Aptitude Battery,” *Psychological Reports*, **67**, 1990, pp. 1219-22.

38. Kranzler JH, “The Construct Validity of the Multidimensional Aptitude Battery: A Word of Caution,” *Journal of Clinical Psychology*, **47**(5), Sep 1991, pp. 691-7.

39. Carretta TR, Retzlaff PD, Callister JD, King RE, “A Comparison of Two U.S. Air Force Pilot Aptitude Tests,” *Aviation, Space, and Environmental Medicine*, **69**(10), Oct 1998, pp. 931-5.

40. Carretta TR, Retzlaff PD, King RE, *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, Armstrong Laboratory, Human Resources Directorate, Mesa, AZ, Dec 1997.

41. Powell DH, Kaplan EF, Whitla D, Weintraub S, Caitlin R, Funkenstein HH, **MicroCog: Assessment of Cognitive Functioning (Version 2.1) Manual**, Psychological Corporation, San Antonio, TX, 1993.

42. Vanderploeg RD, ed., **Clinician's Guide to Neuropsychological Assessment**, 2nd ed., Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, 2000.

43. Chappelle W, Ree MJ, Barto EL, Teachout MS, Thompson WT, *Joint Use of the MAB-II and MicroCog for Improvements in the Clinical and Neuropsychological Screening and Aeromedical Waiver Process of Rated USAF Pilots,* AFRL-SA-BR-TR-2010-0002, U.S. Air Force School of Aerospace Medicine, Brooks City-Base, TX, Jan 2010.

44. Kay GG, **CogScreen-Aeromedical Edition: Professional Manual**, Psychological Assessment Resources, Inc., Odessa, FL, 1995.

LIST OF ABBREVIATIONS AND ACRONYMS

AFOQT	Air Force Officer Qualifying Test
DOR	dropped on request
FlyDef	flying training deficiency
GPA	grade point average
MAB	Multidimensional Aptitude Battery
OTS	Officer Training School
PCSM	Pilot Candidate Selection Method
ROTC	Reserve Officer Training Corps
SD	standard deviation
USAF	United States Air Force
USAFA	U.S. Air Force Academy
USAFSAM	U.S. Air Force School of Aerospace Medicine