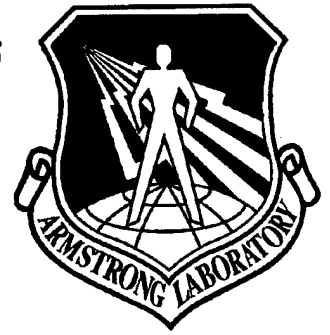


AL/AO-TR-1996-0107



**THE COMPUTERIZED NEUROPSYCHOLOGICAL  
EVALUATION OF US AIR FORCE PILOTS: CLINICAL  
PROCEDURES AND DATA-BASED DECISION**

**Paul D. Retzlaff  
Joseph D. Callister  
Raymond E. King**

**AEROSPACE MEDICINE DIRECTORATE  
CLINICAL SCIENCES DIVISION  
NEUROPSYCHIATRY BRANCH  
2507 Kennedy Circle  
Brooks Air Force Base, TX 78235-5117**

**August 1996**

**Interim Technical Report for Period March 1994 – July 1995**

Approved for public release; distribution is unlimited.

19960926 106

DTIC QUALITY INSPECTED 3

**AIR FORCE MATERIEL COMMAND  
BROOKS AIR FORCE BASE, TEXAS**

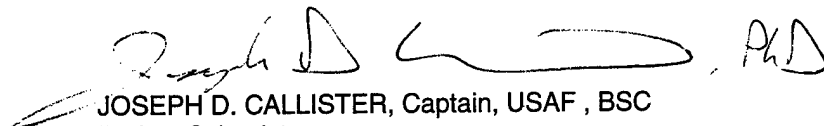
**ARMSTRONG  
LABORATORY**

## NOTICES

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

The Office of Public Affairs has reviewed this technical report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.



JOSEPH D. CALLISTER, Captain, USAF, BSC  
Project Scientist



KENNETH F. BLIFORT, Colonel, USAF, MC, CFS  
Chief, Clinical Sciences Division

# REPORT DOCUMENTATION PAGE

Form Approved  
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE August 1996	3. REPORT TYPE AND DATES COVERED Interim - March 1994 - July 1995	
4. TITLE AND SUBTITLE The Computerized Neuropsychological Evaluation of US Air Force Pilots: Clinical Procedures and Data-Based Decision			5. FUNDING NUMBERS PR - 7350 TA - 32 WU - X1	
6. AUTHOR(S) Paul D. Retzlaff Joseph D. Callister Raymond E. King				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Armstrong Laboratory (AFMC) Aerospace Medicine Directorate Clinical Sciences Division, Neuropsychiatry Branch 2507 Kennedy Circle Brooks Air Force Base, TX 78235-5117			8. PERFORMING ORGANIZATION REPORT NUMBER AL/AO-TR-1996-0107	
9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)			10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES Armstrong Laboratory Technical Monitor: Captain Joseph D. Callister, AL/AOCN, (210) 536-3232.				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words)  The neuropsychological assessment of US Air Force pilots presents several unique problems, given their relatively high cognitive functioning. The United States Air Force currently has a baselining procedure wherein student pilot candidates undergo computerized cognitive assessment. The intent of this assessment is to archive pre-morbid data against which to compare potential future post-accident performance. The current work provides the necessary background, clinical methods and data in order to assess pilots who have suffered cortical insult such as trauma, disease, or exposure to toxin. Methods are delineated for those with pre-morbid testing as well as for those pilots without such testing.				
14. SUBJECT TERMS Aeromedical Evaluation Cognitive Evaluation			15. NUMBER OF PAGES 36	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

## CONTENTS

	Page
SUMMARY.....	1
INTRODUCTION.....	2
Background.....	2
Purpose.....	3
METHOD.....	4
Subjects.....	4
Measures.....	4
Clinical Methods.....	5
RESULTS and APPLICATION.....	7
Change in Performance Method.....	7
Level of Performance Method.....	9
Pattern of Performance Method.....	11
DISCUSSION.....	12
REFERENCES.....	13

## TABLES

### Table Number

1	Means and standard deviations for all MAB variables.....	15
2	Means and standard deviations for CogScreen Speed variables.....	16
3	Means and standard deviations for CogScreen Accuracy variables.....	17
4	Means and standard deviations for CogScreen Throughput variables.....	18
5	Means and Standard Deviations for CogScreen Process variables.....	19
6	Percentiles for MAB variables.....	20
7	Percentiles for Cogscreen Speed variables.....	21
8	Percentiles for Cogscreen Accuracy variables.....	22
9	Percentiles for Cogscreen Throughput variables.....	23
10	Percentiles for Cogscreen Process variables.....	24
11	Percentiles for MAB variable difference scores.....	25

### Appendices

A	CogScreen Variable Definitions.....	26
---	-------------------------------------	----

## PREFACE

This project was funded by Air Force Medical Operating Agency and Armstrong Laboratory.

Appreciation is extended to the technical support staff of the project including SSgt. Pauline M. Etterle, SrA. W. David Taylor and William M. Weaver.

## SUMMARY

The computerized neuropsychological evaluation of US Air Force pilots: Clinical procedures and data-based decisions.

The neuropsychological assessment of US Air Force pilots presents several unique problems given their relatively high cognitive functioning. The United States Air Force currently has a baselining procedure wherein student pilot candidates undergo computerized cognitive assessment. The intent of this assessment is to archive pre-morbid data against which to compare potential future post-accident performance. The current work provides the necessary background, clinical methods, and data in order to assess pilots who have suffered cortical insult such as trauma, disease, or exposure to toxins. Methods are delineated for those with pre-morbid testing as well as for those pilots without such testing.

The computerized neuropsychological evaluation of US Air Force pilots: Clinical procedures and data-based decisions.

## Background

Aviation is one of the most cognitively demanding occupations. Any decline in cognitive ability is of great concern from a number of perspectives. After initial flight training, a number of cognitive insults may result in an occupationally significant cognitive decline. These insults can include chronic alcohol abuse, brain trauma, cerebrovascular insufficiencies, neurodegenerative diseases, and psychiatric disabilities such as depression. The resultant declines in performance may be temporary or permanent. The complexity of aviation jobs and the unforgiving nature of the working environment demands a conservative approach to an occupational return after even the smallest central nervous system insult. At a minimum, medical and neurological evaluations are completed, but in addition, neuropsychological assessment may be indicated.

Regulation, in some instances, is used to guard against the potentially cognitively impaired pilot. Salive (1994), considering the mandatory retirement of commercial pilots over the age of 60, recommended the development of research paradigms and data which might be used in the debate. He discussed the history of the Federal Aviation Administration (FAA) rule as well as the epidemiology of relevant diseases and testing. He concluded that very sensitive and specific tests must be used in the medical and neuropsychological examination of pilots.

Neuropsychology is the examination of brain-behavior relationships. Clinically, it usually takes the form of a psychologist using various "tests" to map the cognitive functions of a patient (Vanderploeg, 1994). These tests assess intelligence, attention, memory, planning and processing, and spatial abilities, as well as other dimensions.

There has been some prior military work in the area of clinical cognitive assessment. Guilmette and Treanor (1986) describe many of the reasons for the importance of the neuropsychological assessment of aviators. They looked at the neuropsychological performance of a group of 15 Army aviators using traditional clinical tests. They found few differences between the aviators and a control group. A potential research confound, however, was the fact that the control group was a remarkably high functioning group of 15 with IQs of 120.

From a civilian perspective, Banich, Stokes, and Elledge (1989) reviewed the literature on the mental status assessment of pilots. They concluded that interview methods lack sensitivity and that existing clinical tests are the alternative of choice. Specifically, they suggested using many of the classic variables found in multiscale intelligence tests. This group later (Stokes, Banich, and Elledge, 1991) demonstrated that

computerized neuropsychological tests were superior to short mental status examinations. To demonstrate this, they compared a group of pilots to a group of cognitively impaired patients.

Currently in the USAF, the Aeromedical Consultation Service of Armstrong Laboratory at Brooks AFB, TX, is responsible for conducting aviator medical, psychiatric, and psychological evaluations whenever a waiver is required to continue flying. Results of the evaluations are provided to the referral source and recommendations as to flying status are made. Standard evaluations include intelligence, neuropsychological, personality, psychopathology, and neurological assessment procedures. As such, most referrals are seen by Neurology, Psychiatry, and Psychology functions.

There are a number of problems in the neuropsychological assessment of any patient and a number specific to aviators. For all patients, it is often difficult to infer pre-morbid levels of functioning. This information, however, is usually needed in order to better assess the current level of functioning and any change from that prior level. From a testing perspective, assessment instruments must be reliable and valid as well as sensitive and specific. All too often tests of limited psychometric value are used. Finally, the relationship between the testing behavior and actual real-life functioning is at best theoretical.

For aviators specifically, methodological difficulties result from the fact that this group possesses atypically high levels of cognitive ability. This often limits the use of certain statistics such as variance, reliability, and validity that are gleaned from much more heterogeneous samples. Aviators perform so well on tests that many assumptions that are used to detect change in patients in general are of limited value when applied to aviators.

In 1994, the USAF began a program to screen pilot training candidates prior to their transfer to Undergraduate Pilot Training bases. The Enhanced Flight Screening (EFS) program included actual flight training in propeller-driven aircraft and medical evaluation. As a part of the medical evaluation, a number of psychological tests were mandated. The primary purpose of the cognitive tests is to archive the individual pilot's scores for future use. The intent is to develop a registry against which future testing might be compared. In essence, pre-morbid data was to be collected on all pilots so that later medical decisions would have an empirical base. As such, the psychological portion of the EFS program (King and Flynn, 1995) includes traditional scales of intelligence as well as newer computerized cognitive tasks.

#### Purpose

The purpose of the present paper is to provide clinical procedures for the evaluation of pilots with cognitive referral



questions and to provide the necessary comparative test norms. Procedures are provided for patients who have pre-morbid EFS testing and for those without such testing.

## METHOD

### Subjects

A sample of 537 Air Force pilot training candidates participated in this study. The sample as a whole had a mean age of 23.5 (standard deviation 4.2) and about 8% were female. Subjects who had been commissioned through Officer Training School, Reserve Officer Training Corps, and the Air National Guard were all college graduates. Approximately, 42% were Juniors at the United States Air Force Academy. Student pilot candidates participated in the baseline cognitive testing during EFS either at the Air Force Academy in Colorado Springs, CO, or at Hondo, TX.

### Measures

The Multidimensional Aptitude Battery (MAB) (Jackson, 1985) is a broad based test of intellectual ability. It was patterned after the Wechsler Adult Intelligence Scale (WAIS-R; correlation = .91), the most widely used individually administered test of intelligence. While the WAIS-R requires about an hour and a half per subject to administer, the MAB can be given to groups and requires about the same amount of total testing time. Additionally, the WAIS-R requires skillful scoring while the MAB has a multiple choice format. All subtests in the WAIS-R have corresponding paper and pencil subtests in the MAB except immediate digit memory. Verbal components tapped include information, comprehension, arithmetic, similarities, and vocabulary. Performance measures include digit symbol coding, picture completion, spatial, picture arrangement, and object assembly. Scores on each of the subtests are scaled to a mean of 50 and a standard deviation of 10. Verbal and performance sub-scores are available as is a full scale intelligence score, each scaled to a mean of 100 and a standard deviation of 15. Reliabilities for the summary scores range from .94 to .98.

Current testing in the USAF Enhanced Flight Screening program (King and Flynn, 1995), other US Air Force research programs (Flynn, Sipes, Grosenbach, and Ellsworth, 1994; Retzlaff and Gibertini, 1987), NASA's astronaut selection procedure, and a number of civilian airline screening procedures include the MAB.

The version of the MAB used in the current study was primarily the computerized version (Retzlaff, King, and Callister, 1995a). Here verbal questions are presented as text on a computer screen and subjects are asked to respond to the computer with an a, b, c, d, or e keyboard entry. The performance items were scanned into computer graphic files and are presented in a window on the monitor. This computerization was done and is used with the consent of the test author with

explicit copyright permission. It is important to note that the 1990 norms for the MAB were used for this study. These norms are used in the computer scoring software from the publisher. Earlier work with the test or other current paper-and-pencil type administrations use the original 1985 norms. Hence, direct comparison with data such as Retzlaff and Gibertini's (1988) may be difficult.

The CogScreen-Aeromedical Edition (Kay, 1995) 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 are generally more fundamental processes such as reaction time. It is not a test of aviation knowledge but considered to include abilities necessary in the performance of aviation duties (Kay and Horst, 1988). There are 11 tasks which result in 65 scores. The tasks include Backward Digit Span (BDS), Math (MATH), Visual Sequence Comparison (VSC), Symbol Digit Coding (SDC), Matching-to-Sample (MTS), Manikin (MAN), Divided Attention (DAT), Auditory Sequence Comparison (ASC), Pathfinder (PF), Shifting Attention (SAT), and Dual Task (DTT). Each of the tasks is usually scored in a number of ways. Typical scorings include task speed, accuracy, and throughput. Throughput is a function of speed and accuracy, basically the number of correct responses per minute. It is indicative of the amount of work accomplished. A number of tasks also include process completion measures which quantify task specific behavior such as control of the computer screen elements. The manual and other research refers to the CogScreen scores by a relatively cryptic variable naming process. These variable names are defined in Appendix A.

The CogScreen is relatively new and represents an attempt by its authors to produce an assessment device which met a number of Federal Aviation Administration requirements. It is currently used in the EFS program by the USAF, by the US Navy, and by a number of commercial airlines. It is published and available from one of the major psychological test publishers.

The CogScreen was used as provided by the test publisher. Computer software administers the test, times the tasks, scores the tests, and archives the data in report form.

### Clinical methods

There are three major manners in which to use the available data (Retzlaff and Gibertini, 1994). The first is the intended purpose of EFS. This procedure compares the archived data (pre-morbid) to later testing (post-morbid), presumably after some sort of cognitive insult.

The other two procedures acknowledge the fact that not all pilots will have archived pre-morbid data. This may be the case because either they became pilots before the program began or they become pilots after the program was terminated (if indeed the program is terminated). These two procedures use data

developed from those taking the EFS testing. As such, the second procedure looks at the relative ability level of the new patient given the known ability levels for the tested group. The third and final method uses a number of the tests for a new subject as control conditions for other tests taken at the same time.

Change in Performance Method The first method is a pre-test, post-test paradigm. It is the most reliable but requires prior, pre-morbid testing data against which to compare later testing. In the general clinical case, a patient may have prior intelligence and neuropsychological testing, been exposed to some cortical insult, and then re-tested. An example might be a patient in the Veteran's Administration system. It would be common for a patient to have a prior intelligence test such as a WAIS-R somewhere in the system, have some sort of cortical insult such as a stroke or head injury, and then be re-tested on the same intelligence test. Here the results of the first testing can be used as a reference for the second testing. A significant decrement across testings would establish the existence of a dementia and gauge the general severity of it.

The degree to which test scores may vary from one testing to the next can be established statistically. "Normal" or chance degrees of differences can be established through studying the stability of normal subjects across two testing periods. The first testing is correlated with the second to establish a stability (reliability) coefficient. This coefficient can be used to determine a confidence band around a score. Performance beyond this confidence band would suggest performance decrements beyond what might be expected by chance.

For aviators who have participated in the EFS program, pre-morbid data is available and can be retrieved from Armstrong Laboratory. Knowing the aviator's initial performance, the stability coefficient of the test, and the variability of the test for aviators, confidence bands can be established for an individual aviator. Performance below what can be expected statistically on the MAB or CogScreen may be taken as evidence of an impairment.

Level of Performance Method To date, only a very small percentage of USAF aviators have archived EFS testing. As such, methodologies are necessary for the assessment of aviators without pre-morbid testing. Here the EFS data on MAB and CogScreen variables may be used as a group reference. Pilots with poor performance on testing following some insult may be inferred to be at that low level of performance due to the cortical insult. Aviators who are found to be in the bottom one percent following some trauma, for example, are statistically more likely to be at that level due to the trauma than due to their initial performance. In other words, there would only be a one percent chance that the aviator was pre-morbidly at that low level of performance.

In order to effectively utilize this approach, a number of statistics and tables are necessary. First, the means and standard deviations of a large sample of fairly similar individuals is required. This provides the norm against which to compare a new individual's scores. In addition to these statistics, percentile levels of various scores are often of use. While the mean and standard deviations model the underlying distribution of test scores when the distribution is normal, they do not model skewed distributions well when there is an asymmetry in scores. Providing the scores of a distribution at critical percentile points allows the scores of new patients to be very accurately placed relative to their peers.

Pattern of Performance Method While the above method uses a large group of subjects as the comparison for an individual's post-insult scores, it is also possible to use some elements of the person's own performance to make conclusion regarding cognitive change. A common approach uses the effects of aging on various types of test performance as a model. It has long been known that some types of intellectual ability are fairly sensitive to aging and other types are quite resistant to change. Classically, these are referred to as "hold" and "don't hold" variables. Scores on tasks such as vocabulary and general information generally are similar across age brackets. These tasks tend to "hold" as one ages. Scores on other tasks such as performance type tests like speed dependent visuomotor ability usually drop off with age. Here, somewhere in the fifth decade of life, performances "don't hold" and begin a fairly constant decline.

Applying this method to younger patients who have had some type of cortical insult suggests that larger differences in scores between "hold" and "don't hold" tests is associated with greater levels of impairment. It is common, for example, to look at the difference between the Vocabulary subtest on the WAIS-R and the Digit Symbol subtest. If the Digit Symbol subtest is more than 2 or 3 standard scores below the Vocabulary subtest score (and there is history of insult), there is a good likelihood of impairment.

There are always naturally occurring differences between two sub-tests on any test. It is, therefore, necessary to quantify this natural difference so that referred aviators might be compared to the "normal" differences. Aviators whose difference scores between two tests are in the top 99% of non-impaired aviators can be assumed to have that level of difference due to insult, as the a priori chance of that difference is quite low.

## Results and Application

### Change in Performance Method

Table 1 provides the means and standard deviations for each of the MAB scores. These include summary scores as well as scaled and raw scores. The scaled scores are based upon the 1990

norms. The raw scores are provided here and in subsequent tables in the event that there is a re-norming of the test. As can be seen, pilots are on average quite intelligent with Full Scale IQ scores of 119. This table also includes the stability coefficient, the standard error of measurement, and the 95% confidence band for each of the scores. The stability coefficient is based upon the testing and retesting of a group of subjects during the development of the test. It indicates the degree to which scores remain constant across time. The standard error of estimate statistic indicates the variability of scores that could be expected from multiple testings of the same person. Finally, the 95% confidence band indicates the differences in scores that might be expected at the 95% probability level. This final confidence band can be applied to any individual's scores. If a second testing is below the confidence band, the performance should be interpreted as lower and more deficient than what can be expected simply due to measurement error.

As an example, suppose a pilot received a Full Scale IQ score of 125 during initial EFS screening. The pilot is then involved in a car accident with a brief coma. The pilot is referred for follow-up cognitive testing. The expected range of scores for this pilot would be 125 plus or minus 2.38 points. As such, the range would be 123 to 127. The MAB is re-administered and the Full Scale IQ score is 118. Since this is well below the bottom of the confidence band (123), there is good reason to suspect a true decrement in ability. Obviously, it is another question whether an IQ of 118 is too low to continue flying; nevertheless, an impairment is verified. Further testing and other evidence can assess the question of continued flying.

There are ten subscales which can also be used to answer more specific functional questions in the same manner. It is of particular importance when a referral question specifically mentions an error of concern such as spatial ability and subsequent testing indicates performance on the spatial subtest well below the confidence band. Additional evidence might be gathered from the number of subscales below the bands. A pilot with only one of the tests below the band is very different from a pilot with all ten subtests below the bands.

With 65 variables, the CogScreen is somewhat difficult to interpret (See Appendix A for variable names). In order to better understand the data, it is presented not by subtest but by type of score. As such, speed variables are presented first, followed by accuracy, throughput, and process variables.

Table 2 provides not only the means and standard deviations for the CogScreen speed variables but also the stability coefficient, the standard error of estimate, and the 95% confidence band. The stability coefficient was taken from the test manual and used specifically to develop the other two statistics for this sample.

Here, for example, a subject's reaction time speed score on the Math task would have to be banded by plus and minus 8.26 seconds. As such, a subject with a pre-morbid score of 30.00 seconds would have a 95% statistical probability of producing a score between 21.74 and 38.26 seconds. A clinically important finding would be a score significantly slower such as 42 seconds. In this example, a pilot with a pre-morbid score of 30 seconds probably has a decline from prior functioning with that score of 42 seconds. Conversely, a post-morbid score of 35 seconds is within the measurement error range and should not be clinically interpreted as a decline.

With so many speed scores, it is important not to calculate so many statistics on a single patient that the method becomes a "fishing trip" with a "drift net". The two CogScreen tasks with the best speed characteristics are probably the MTS (Matching to Sample) and MAN (Manikin). These tasks require a small amount of cognitive performance directed toward a fairly focal stimuli. With average performance in the one and a half to two second range, there is sufficient room for variable performance. Tasks which have much shorter reaction times are probably prone to be confounded by the use of the light pen, the use of large muscle groups, subtle shifts of position, administration differences, and software changes. Tasks such as MATH are not true reaction times. The 30 seconds or so of task time includes attention, reading speed, math calculation time, and reaction time. As such, it is a heterogeneous task, and hence of limited interpretive value here.

Tables 3, 4, and 5 provide the means and standard deviations for the CogScreen accuracy, throughput, and process variables. The accuracy scores have so little variance in normal pilots that the calculation of stability coefficients, standard errors of measurement/ estimate, and confidence bands is inappropriate. This lack of variance is also noted in the manual for the normative sample. The reason that the scores vary so little is due to "ceiling effect". The tasks are so easy that most subjects (at times over 90%) get all tasks correct and as such there is no separation of performance on the high end of ability. Since throughput variables are the product of speed and accuracy variables, they add little information over the speed data. Finally, the manual does not present stability data for the process variables and as such confidence bands cannot be calculated. Here is an example of where a USAF stability study would allow for such data.

#### Level of Performance Method

Table 6 provides the percentile levels for the MAB variable distributions. A subject with a score of 129 would be at the 95% and be quite intelligent compared to other pilots. For clinical purposes with a patient who did not have prior testing, these data can be interpreted as the probability of a post-insult decrement in functioning.

The chances that a pilot has a Full Scale IQ score of 100 is about 1%, because only 1% of the sample have Full Scale IQ scores of 100 or less. One way to interpret this data clinically is to say that there is a 99% chance that the pilot with the IQ of 100 had an IQ of greater than 100 prior to any cognitive insult. Here the very fact of exceptionally low performance is in and of itself unlikely and most probably due to clinical factors.

In general, scores in the lower 1% and 5% levels are probably clinically relevant. Again, the quality (which scales) and quantity (how many scales) are of interest. Performance scores and tasks are more important for aviators and also more prone to cognitive decline with insult. Conversely, pilots with scores in the top 95% and 99% are probably able to return to duty and clinical significant impact is highly unlikely.

Tables 7, 8, 9, and 10 provide similar cutscores for the CogScreen speed, accuracy, throughput, and process distributions. For the speed data in Table 7, performance is in seconds and therefore larger numbers represent poorer performance. While on the MAB higher scores are better, here lower scores are better. Very fast answering of the Math items might result in a score of 15 seconds. This would place that subject at the 5% level, a very good performance.

A patient, however, who spends 45 seconds on average would be somewhere between the 95% and 99% level. That patient had a very small chance of taking that much time given the group norms and so is probably impaired. Again, the quality and quantity of scores must be part of the clinical decision process.

Again, as with the speed variables in the CogScreen, it is recommended that the Matching to Sample (MTS) and Manikin (MAN) tasks be used for most clinical work. They exhibit good range across the sample and are less prone to error than the faster, pure reaction time tasks.

Table 8 provides the tail of the distribution associated with low accuracy scores. Full tables are not possible due to the limited variance of these scores. In essence, most pilots got these tasks right with a few pilots getting some tasks wrong. Using MATH as the example again, a pilot who only gets a .20 proportion of the MATH questions correct is at the bottom 1% of the distribution. A .40 proportion would place that pilot at only the 15% level. Either score should be of clinical concern.

Table 9 presents the throughput data. Here, higher scores represent very fast, accurate, and efficient cognitive processes. Low scores represent poor performance. A throughput of 0.3 on the MATH task would represent a performance at the first percentile of the distribution. This would suggest a impairment relative to the norms.

Finally, Table 10 presents the distributions for the process variables. The table's footnote indicates the direction of performance and the tails of clinical concern. Here, again, a number of the variables had highly skewed distributions with limited variance and only a limited number of distribution points could be mapped.

#### Pattern of Performance Method

Table 11 provides the statistically expected differences in scaled scores across tests given to a single subject at a single point in time. The MAB is used here because the variables are widely used and understood. The CogScreen is not presented because no theory or research exists on its interscale behavior in impaired individuals.

The approach here is that variables such as Vocabulary and Information are relatively resistant to cognitive insult. The performance tasks (Digit Symbol, Picture Completion, Spatial, Picture Arrangement, and Object Assembly) are far more likely to be affected by an impairing incident. Difference scores, however, will naturally vary quite widely in non-impaired individuals and must be modeled.

To develop this data, the scaled scores for each of the performance tasks was subtracted from the scaled score of Vocabulary and Information. This resulted in a distribution of difference scores for the sample. The means and standard deviations are presented in Table 11. On average, pilots have better performance scores than Vocabulary scores as evidenced by the negative difference scores. Their scores on Information are more similar to, and slightly better than, their scores on the performance tasks with difference scores of generally 1 to 3 points.

The data of interest are those differences which are positive and large. This would clinically suggest that performance type ability is well below the traditional "hold" verbal tests. The "hold" tests would have "held" and the "don't hold" tests would have "not held". The bottom line of a positive and large score would be a cognitive impairment.

If a patient had a scaled Vocabulary score of 60 and a Digit Symbol score of 45, the difference would be 15 points. Looking at the table, a 15 point difference would place this patient well above the 99th percentile. A clinician could be 99% certain that such scores would not be found in non-impaired pilots.

It is recommended that the Scaled Vocabulary score minus the Scaled Digit Symbol score be used for most purposes. Vocabulary seems to behave best in this population and appears to have the most stable norming across studies using the MAB. Digit Symbol is a complex, heterogeneous task which is sensitive to many functional declines. The raw score difference scores are



unstable due to the lack of a common underlying metric and are provided here for reference only.

## DISCUSSION

The accurate assessment of the cognitive functioning of pilots is essential. The lives and careers of pilots and the lives of crews and passengers may depend upon it. The USAF also is interested in increasing mission effectiveness, reducing training costs, and managing retention.

The USAF EFS program provides an opportunity to collect large sets of cognitive data on pilot candidates (Callister, King, and Retzlaff, 1995). No other study or function has ever allowed for such large samples or for the archiving of individual data.

Three clinical methods for the neuropsychological assessment of pilots have been delineated. A method using pre-morbid test data for those pilots with archived EFS data has been explored. Additionally, two methods have been explained for the testing of pilots without pre-morbid testing available. The necessary statistical tables are presented for clinical use.

A number of caveats must be mentioned. First, these data are from pilot candidates. As such there is some chance that the data are not as precise as they might be. A number of studies, however, have found very similar intelligence test data. Also Retzlaff, King, and Callister (1995b) found no differences in intelligence between those entering pilot training and those finishing. The CogScreen is less well known and larger differences may operate.

It would also have been better to use stability coefficients which had been calculated from an Air Force pilot sample. The use of general stability coefficients from the test manuals are within the normal range of practice, but a one year test-retest study of a group of mid-career pilots would have provided much more specific statistics.

Finally, it is important to note that this is a relatively atypical approach to neuropsychology driven by the unique needs of the USAF medical baselining requirements. Psychology has a long history of neuropsychological tests, assessment, and methods. Traditional neuropsychological assessment includes many tests across many hours of individualized testing. It is fully expected that the current work will be in addition to, not in place of, the traditional techniques.

## REFERENCES

- Banich, M. T., Stokes, A., & Elledge, V. C. (1989). Neuropsychological screening of aviators: A review. Aviation, Space, and Environmental Medicine, 60, 361-366.
- Callister, J. D., King, R. E., & Retzlaff, P. (1995). Cognitive assessment of USAF pilot training candidates: Multidimensional Aptitude Battery and CogScreen Aeromedical Edition. (AL/AO-TR-1995-0125). Brooks AFB: Armstrong Laboratory.
- Flynn, C. F., Sipes, W. E., Grosenbach, M. J., and Ellsworth, J. (1994). Top performer survey: Computerized psychological assessment of aircrew. Aviation, Space, and Environmental Medicine, 65, 39-44.
- Guilmette, T. J. & Treanor, J. J. (1986). Baseline and comparative neuropsychological data on U. S. Army aviators. Aviation, Space, and Environmental Medicine, 57, 950-953.
- Jackson, D. N. (1985). Multidimensional Aptitude Battery. Port Huron, MI: Research Psychologists Press.
- Kay, G. G. (1995). CogScreen-Aeromedical Edition Professional Manual. Odessa, FL: Psychological Assessment Resources, Inc.
- Kay, G. G., & Horst, R. L. (1988). Methods for evaluating cognitive function: A review of mental status tests, neuropsychological procedures, and performance-based approaches. Technical Report submitted to the Federal Aviation Administration, Civil Aeronautical Medical Institute, [Contract No. DTFA-02-87-87069], Oklahoma City, OK.
- King, R. E. and Flynn, C. F. (1995). Defining and measuring the "Right Stuff": Neuropsychiatrically Enhanced Flight Screening (N-EFS). Aviation, Space, and Environmental Medicine, 66, 951-956.
- Retzlaff, P. and Gibertini M. (1987). Air Force pilot personality: Hard data on "The Right Stuff". Multivariate Behavioral Research, 22, 383-399.
- Retzlaff, P. and Gibertini, M. (1988). The objective psychological testing of Air Force officers in pilot training. Aviation, Space, and Environmental Medicine, 59, 661-663.
- Retzlaff, P. & Gibertini, M. (1994). Neuropsychometric issues and problems. In Vanderploeg, R. (ed.), Clinician's Guide to Neuropsychological Assessment. Hillsdale, NJ: Erlbaum.
- Retzlaff, P., King, R. E., & Callister, J. D. (1995a). Comparison of a computerized version to a paper/ pencil version of the Multidimensional Aptitude Battery. (AL/AO-TR-1995-0121). Brooks AFB: Armstrong Laboratory.

- Retzlaff, P., King, R. E., & Callister, J. D. (1995b). US Air Force pilot training completion and retention: A ten year follow-up on psychological testing. (AL/AO-TR-1995-0124). Brooks AFB: Armstrong Laboratory.
- Salive, M. E. (1994). Evaluation of aging pilots: Evidence, policy, and future directions. Military Medicine, 159, 83-86.
- Stokes, A. F., Banich, M. T., and Elledge, V. C. (1991). Testing the tests- An empirical evaluation of screening tests for the detection of cognitive impairment in aviators. Aviation, Space, and Environmental Medicine, 62, 783-788.
- Vanderploeg, R. (ed.) (1994). Clinician's Guide to Neuropsychological Assessment. Hillsdale, NJ: Erlbaum.

Table 1

Means and standard deviations for all MAB variables.

Variable	Mean	SD	r	SEE	95%
<u>IQ Scores</u>					
Full Scale	119.3	7.0	.97	1.21	2.38
Verbal	118.1	7.0	.95	1.57	3.07
Performance	118.0	8.8	.96	1.76	3.45
<u>Scaled Scores</u>					
Information	67.4	6.8	.97	1.18	2.31
Comprehension	60.1	4.1	.95	0.92	1.80
Arithmetic	62.3	6.5	.88	2.25	4.41
Similarities	62.1	4.8	.83	1.98	3.88
Vocabulary	58.4	6.4	.90	2.02	3.97
Digit Symbol	66.4	6.8	.90	2.15	4.21
Picture Completion	63.7	6.8	.94	1.67	3.26
Spatial	63.5	7.3	.93	1.93	3.79
Picture Arrangement	60.1	7.2	.87	2.60	5.09
Object Assembly	64.5	7.5	.93	1.98	3.89
<u>Raw Scores</u>					
Information	29.4	4.5	.97	0.78	1.53
Comprehension	23.4	2.2	.95	0.49	0.96
Arithmetic	15.7	2.0	.88	0.69	1.36
Similarities	27.8	3.0	.83	1.24	2.42
Vocabulary	29.2	5.7	.90	1.80	3.53
Digit Symbol	29.2	3.4	.90	1.08	2.11
Picture Completion	26.9	3.7	.94	0.91	1.78
Spatial	36.8	6.8	.93	1.80	3.53
Picture Arrangement	12.6	2.1	.87	0.76	1.48
Object Assembly	15.7	3.1	.93	0.82	1.61

Note: N=537, r is the stability (reliability) taken from the MAB manual, SEE is the Standard Error of Estimate, and 95% is the 95% confidence interval.

Table 2

Means and standard deviations for CogScreen Speed variables.

Variable	Mean	SD	r	SEE	95%
4 MATHRTC	27.25	8.79	.77	4.22	8.26
7 VSCRTC	2.24	.51	.89	.17	.33
14 MTSRTC	1.47	.28	.79	.13	.25
17 MANRTC	1.98	.38	.85	.15	.29
19 DATIRTC	.40	.07	.68	.04	.08
21 DATDRTC	.69	.20	.63	.12	.24
23 DATSCRTC	2.15	.53	.84	.21	.42
27 ASCRTC	.98	.24	.75	.12	.24
30 PFNRTC	.85	.16	.77	.08	.15
34 PFLRTC	.79	.13	.75	.07	.13
38 PFCRTC	1.20	.30	.80	.13	.26
42 SATADRTC	.70	.10	.80	.04	.09
45 SATABRTC	.68	.09	.72	.05	.09
48 SATINRTC	.86	.15	.91	.05	.09
51 SATDIRTC	.95	.21	.76	.10	.20
57 DTTAABS	24.12	19.50	.65	11.54	22.61
59 DTTDABS	49.42	26.06	.85	10.09	19.78
61 DTTPARTC	.48	.19	.76	.09	.18
64 DTTDRTC	.66	.24	.72	.13	.25

Note: N=512, r is the stability (reliability) taken from the CogScreen manual, SEE is the Standard Error of Estimate, and 95% is the 95% confidence interval. All scores are in seconds except DTTAABS and DTTDABS which are distance measures in fixed seconds. Appendix A provides full variable definitions.

Table 3

Means and standard deviations for CogScreen Accuracy variables.

Variable

	Mean	SD
2 BDSACC	.89	.12
3 MATHACC	.72	.19
6 VSCACC	.97	.03
9 SDCACC	.99	.01
11 SDCIRACC	.94	.13
12 SDCDRACC	.93	.15
13 MTSACC	.95	.05
16 MANACC	.93	.09
24 DATSCACC	.89	.07
26 ASCACC	.90	.10
29 PFNACC	.99	.01
33 PFLACC	.99	.01
37 PFCACC	.98	.03
41 SATADACC	.98	.03
44 SATAACC	.99	.03
47 SATINACC	.97	.03
50 SATDIACC	.67	.11
60 DTTPAACC	.93	.07
63 DTTPDACC	.86	.11

Table 4

Means and standard deviations for CogScreen Throughput variables.

Variable

	Mean	SD
5 MATHPUT	1.82	1.22
8 VSCPUT	27.56	6.20
10 SDCPUT	33.74	6.00
15 MTSPUT	40.44	7.73
18 MANPUT	29.51	7.05
25 DATSCPUT	26.32	6.47
28 ASCPUT	58.79	17.48
31 PFNPUT	72.00	12.86
35 PFLPUT	77.46	12.24
39 PFCPUT	51.83	12.54
43 SATADPUT	86.55	12.77
46 SATACPUT	88.51	11.20
49 SATINPUT	69.59	11.64
52 SATDIPUT	44.64	11.68
62 DTTPAPUT	131.25	46.15
65 DTTPDPUT	90.85	38.48

Table 5

Means and Standard Deviations for CogScreen Process variables.

Variable			
-----			
		Mean	SD
-----			
20	DATIPRE	2.52	1.80
22	DATDPRE	2.22	2.04
32	PFNCOOR	0.80	0.33
36	PFLCOOR	0.95	0.35
40	PFCCOOR	0.87	0.31
53	SATDIRUL	6.96	2.50
54	SATDIFAI	2.15	1.92
55	SATDIPER	1.89	2.51
56	SATDINON	1.57	2.71
58	DTTAHIT	0.92	1.95
59	DTTDHIT	3.49	3.39
-----			



Table 6

Percentiles for MAB variables.

Variable	1%	5%	15%	50%	85%	95%	99%
<u>IQ Scores</u>							
Full Scale	100	107	112	119	126	129	133
Verbal	99	106	110	118	125	129	132
Performance	93	101	109	118	126	131	135
<u>Scaled Scores</u>							
Information	47	53	59	68	73	76	79
Comprehension	49	52	56	60	64	65	67
Arithmetic	45	51	54	60	67	70	79
Similarities	46	53	56	62	66	69	70
Vocabulary	43	47	52	58	65	68	73
Digit Symbol	48	54	58	66	72	76	78
Picture Completion	45	51	56	64	69	73	75
Spatial	37	52	57	63	70	74	77
Picture Arrangement	42	48	52	59	65	72	76
Object Assembly	40	49	56	65	70	72	74
<u>Raw Scores</u>							
Information	16	20	24	30	33	35	37
Comprehension	17	19	21	23	25	26	27
Arithmetic	10	12	13	15	17	18	21
Similarities	18	22	24	28	30	32	33
Vocabulary	15	19	23	29	35	38	42
Digit Symbol	20	22	25	29	32	34	35
Picture Completion	17	20	23	27	30	32	33
Spatial	12	26	31	36	43	47	49
Picture Arrangement	7	9	10	12	14	16	17
Object Assembly	5	9	12	16	18	19	20

Table 7

Percentiles for Cogscreen Speed variables.

Variable	1%	5%	15%	50%	85%	95%	99%
4 MATHRTC	6.62	15.00	18.39	26.20	36.07	44.01	49.71
7 VSCRTC	1.30	1.48	1.75	2.18	2.67	3.18	3.91
14 MTSRTC	0.92	1.08	1.20	1.44	1.72	1.92	2.21
17 MANRTC	1.26	1.41	1.57	1.95	2.39	2.68	2.99
19 DATIRTC	0.25	0.28	0.33	0.40	0.46	0.52	0.59
21 DATDRTC	0.36	0.44	0.52	0.66	0.87	1.09	1.34
23 DATSCRTC	1.36	1.48	1.63	2.05	2.66	3.15	3.76
27 ASCRTC	0.43	0.62	0.74	0.97	1.22	1.39	1.73
30 PFNRTC	0.55	0.63	0.70	0.83	1.01	1.16	1.34
34 PFLRTC	0.55	0.61	0.66	0.77	0.91	1.04	1.15
38 PFCRTC	0.66	0.82	0.92	1.15	1.48	1.75	2.20
42 SATADRTC	0.51	0.56	0.60	0.68	0.80	0.90	1.02
45 SATACRT	0.51	0.55	0.59	0.67	0.77	0.85	0.98
48 SATINRTC	0.61	0.67	0.71	0.84	1.01	1.16	1.31
51 SATDIRTC	0.61	0.71	0.77	0.92	1.12	1.31	1.77
57 DTTAABS	3.27	4.00	6.13	18.17	43.33	65.75	86.56
59 DTTDABS	9.44	12.93	18.32	47.37	79.75	94.12	104.28
61 DTTPARTC	0.21	0.28	0.32	0.43	0.68	0.87	1.13
64 DTTDRTC	0.25	0.34	0.43	0.61	0.88	1.13	1.44

Table 8

Percentiles for Cogscreen Accuracy variables.

Variable	5%	15%
2 BDSACC	.50	.67
3 MATHACC	.20	.40
6 VSCACC	.85	.90
9 SDCACC	.95	.97
11 SDCIRACC	.50	.83
12 SDCDRACC	.33	.50
13 MTSACC	.80	.85
16 MANACC	.70	.80
24 DATSCACC	.74	.80
26 ASCACC	.60	.70
29 PFNACC	.96	n/a
33 PFLACC	.92	.96
37 PFCACC	.88	.92
41 SATADACC	.83	.92
44 SATAACC	.83	.92
47 SATINACC	.88	.91
50 SATDIACC	.42	.55
60 DTTPAACC	.77	.86
63 DTTPDACC	.61	.74

Note: Performance in general was so high on the PFNACC task that there was insufficient variability to derive a 15% cutscore.

Table 9

Percentiles for Cogscreen Throughput variables.

Variable	1%	5%	15%	50%	85%	95%	99%
5 MATHPUT	0.3	0.7	1.0	1.7	2.5	3.2	5.8
8 VSCPUT	14	19	22	27	34	40	45
10 SDCPUT	22	25	28	33	39	45	51
15 MTSPUT	23	29	33	40	46	53	62
18 MANPUT	14	18	22	29	37	41	46
25 DATSCPUT	13	16	20	26	33	37	41
28 ASCPUT	25	35	44	57	73	90	121
31 PFNPUT	44	51	59	71	85	95	105
35 PFLPUT	51	57	66	77	91	97	109
39 PFCPUT	25	33	40	51	64	72	84
43 SATADPUT	57	64	73	88	99	105	115
46 SATACPUT	59	69	77	89	99	106	112
49 SATINPUT	42	50	58	69	82	87	97
52 SATDIPUT	16	21	32	45	56	62	67
62 DTTAPUT	36	56	77	131	175	202	251
65 DTTDPUT	20	37	53	86	126	157	216

Table 10

Percentiles for Cogscreen Process variables.

Variable	5%	15%	50%	85%	95%
32 PFNCOOR	0.20	0.40	0.80	1.10	1.30
36 PFLCOOR	0.30	0.60	0.90	1.30	1.50
40 PFCCOOR	0.40	0.50	0.80	1.10	1.50
53 SATDIRUL	0	4			
20 DATIPRE				4	6
22 DATDPRE				4	6
54 SATDIFAI				4	6
55 SATDIPER				4	6
56 SATDINON				3	7
58 DTTAHIT				2	5
59 DTTDHIT				8	10
N	512				

Note: Only the Coordination variables had sufficient range and resolution to allow percentiles across all ranges. Lower scores on SATDIRUL indicates poorer performance. Higher scores on all other variables indicates poorer performance.

Table 11

Percentiles for MAB variable difference scores.

Variable	Mean	SD	85%	95%	99%
-----					
Scaled Vocabulary minus Scaled:					
-----					
Digit Symbol	-8.0	8.6	0	6	11
Picture Completion	-5.3	7.7	2	8	15
Spatial	-5.1	8.9	3	9	19
Picture Arrangement	-2.4	8.9	6	12	17
Object Assembly	-6.1	8.9	2	7	17
Scaled Information minus Scaled:					
-----					
Digit Symbol	0.9	9.1	10	15	22
Picture Completion	3.7	7.9	11	17	23
Spatial	3.4	8.8	12	18	26
Picture Arrangement	6.6	8.6	15	19	25
Object Assembly	2.9	9.0	11	17	22
Raw Vocabulary minus Raw:					
-----					
Digit Symbol	-0.1	6.2	6	10	14
Picture Completion	2.2	5.7	8	11	16
Spatial	-7.7	8.1	0	5	14
Picture Arrangement	16.5	5.8	22	25	30
Object Assembly	13.4	6.0	19	23	26
Raw Information minus Raw:					
-----					
Digit Symbol	0.2	5.4	5	8	12
Picture Completion	2.5	4.8	7	10	14
Spatial	-7.4	7.2	-1	4	13
Picture Arrangement	16.8	4.5	21	23	25
Object Assembly	13.7	4.9	18	21	23
-----					

## Appendix A

### CogScreen Variable Definitions

#### Backward Digit Span

1 BDSACC Accuracy

#### Math

2 MATHACC Accuracy

3 MATHRTC Speed

4 MATHPUT Thruput

#### Visual Sequence Comparison

5 VSCACC Accuracy

6 VSCRTC Speed

7 VSCPUT Thruput

#### Symbol Digit Coding

8 SDCACC Accuracy

9 SDCPUT Thruput

10 SDCIRACC Immediate Recall Accuracy

11 SDCDRACC Delayed Recall Accuracy

#### Matching to Sample

12 MTSACC Accuracy

13 MTSRTC Speed

14 MTSPUT Thruput

#### Manikin Test

15 MANACC Accuracy

16 MANRTC Speed

17 MANPUT Thruput

#### Divided Attention Test

18 DATIRTC Indicator alone speed

19 DATIPRE Indicator alone premature response

20 DATDRTC Indicator dual speed

21 DATDPRE Indicator dual premature response

22 DATSCACC Sequence comparison accuracy

23 DATSCRTC Sequence comparison speed

24 DATSCPUT Sequence comparison thruput

#### Auditory Sequence Comparison

25 ASCACC Accuracy

26 ASCRTC Speed

27 ASCPUT Thruput

Pathfinder

28 PFNACC	Number accuracy
29 PFNRTC	Number speed
30 PFNPUT	Number thruput
31 PFNCOOR	Number coordination
32 PFLACC	Letter accuracy
33 PFLRTC	Letter speed
34 PFLPUT	Letter thruput
35 PFLCOOR	Letter coordination
36 PFCACC	Combined accuracy
37 PFCRTC	Combined speed
38 PFCPUT	Combined thruput
39 PFCCOOR	Combined coordination

Shifting Attention Test

40 SATADACC	Arrow direction accuracy
41 SATADRTC	Arrow direction speed
42 SATADPUT	Arrow direction thruput
43 SATAACACC	Arrow color accuracy
44 SATAACRTC	Arrow color speed
45 SATAACPUT	Arrow color thruput
46 SATINACC	Instruction accuracy
47 SATINRTC	Instruction speed
48 SATINPUT	Instruction thruput
49 SATDIACC	Discovery accuracy
50 SATDIRTC	Discovery speed
51 SATDIPUT	Discovery thruput
52 SATDIRUL	Discovery rule shifts completed
53 SATDIFAI	Discovery failed set
54 SATDIPER	Discovery perseveration errors
55 SATDINON	Discovery nonconcept response

Dual Task Test

56 DTTAABS	Tracking alone error
57 DTTAHIT	Tracking alone boundary hits
58 DTTDABS	Tracking dual error
59 DTTDHIT	Tracking dual boundary hits
60 DTTPAACC	Previous number alone accuracy
61 DTTPARTC	Previous number alone speed
62 DTTPAPUT	Previous number alone thruput
63 DTTPDACC	Previous number dual accuracy
64 DTTDRTC	Previous number dual speed
65 DTTDPPUT	Previous number dual thruput