A Systematic Review of Neuropsychological Tests Used to Assess Pilots for Fitness to Fly

Prepared by:
Pacific Northwest Evidence-based Practice Center
Oregon Health & Science University
Mail Code: BICC
3181 SW Sam Jackson Park Road
Portland, OR 97239
www.ohsu.edu/epc

Investigators:
Annette M. Totten, Ph.D.
Eilis Boudreau, M.D., Ph.D.
Tamara P. Cheney, M.D.
Cynthia Davis-O'Reilly, B.S.

Acknowledgements
We would like to thank the following people at the Pacific Nortwest Evidence Based Practice Center at Oregon Health & Science University for their assistance in producing this document Elaine Graham, MLS and Tracy Dana MLS. We also like to thank Pam Day from the Aerospace Medical Association for helping us locate conference presentations.

In designing the study questions and methodology at the outset and during the preparation of this report, the Pacific Northwest Evidence-based Practice Center (EPC) consulted an Advisory Panel. This panel included members representing technical and content experts and potential end users of the research. Broad expertise and perspectives were sought. Divergent and conflicting opinions are common and perceived as healthy scientific discourse that results in a thoughtful, relevant systematic review. Advisory Panel members were not involved in the analysis of the evidence or the writing of the report. After the report was drafted, it was sent to experts for review and comment. Reviewers comments were considered, but ultimately the research team decided on the final content of this report. Therefore, study questions, design, methodological approaches, and/or conclusions do not necessarily represent the views of individual Advisory Panel members or Reviewers.

Advisory Panel
Fred Bonato, PhD, Aerospace Medical Association
Nick Caplan, PhD, Aerospace Medicine Systematic Review Group
James Devoll, MD, Office of Aerospace Medicine, Federal Aviation Administration
Jay Dorothy, Allied Pilots Association
John Hastings, MD, Neurologist
Pete Lewis, Allied Pilots Association
Muriel Lezak, PhD, Neuropsychologist Emeritus Professor Oregon Health and Science University
Ed Miles, PhD, Clinical Psychologist Allied Pilots Association
Mona Nasser, DDS Cochrane Methods Group
Scott Rossow, D.O. CFII, Aerospace Medical Certification Division, Federal Aviation Administration
Andrew Winnard, PhD, Aerospace Medicine Systematic Review Group

Reviewers
Robert W. Elliott, PhD, ABCN, ABN, Aviation Neuropsychologist
Randy Georgemiller, PhD, ABPP, Neuropsychologist, Federal Aviation Administration
Gary G. Kay, PhD, ABN, ABAP, Senior Neuropsychology Consultant to US Federal Air Surgeon
Disclosure: Dr. Kay is the author and publisher of CogScreen LLC. Dr. Kay is the author of Aeromedical Psychology
Joshua R. Potocko, MD, MPH, Occupational Medicine Specialist, US Navy
Max Trenerry, PhD, ABPP-Cn ANt, Professor of Psychology and Consultant, Mayo Clinic, Rochester MN
Alex Wolbrink, MD, MS, Aerospace Medicine Specialist, Monument, Colorado

Advisory Panel members and Reviewers were requested to disclose any potential financial conflicts of interest and any other relevant business or professional conflicts of interests. Dr. Treneery is the author of two neuorpsycchological tests published by PAR. Dr. Kay is the developer of the Cog-Screen. No other conflicts were disclosed

Funding
This project was funded in part by contributions from the Allied Pilots Association, the Aerospace Medicine Association, and private donors; funds were administered through the OHSU Foundation. The authors of this report are responsible for its content. Funders did not directly participate in the literature search, determination of study eligibility criteria, data analysis or interpretation, or preparation, review, or approval of the report. Statements in the report do not necessarily represent the official views of or imply endorsement by the funders.

Table of Contents

Introduction

Background

Aviation safety is achieved through the operation of safe aircraft, in safe airspace, by a safe pilot. Optimal neuropsychological functioning is critical for safely piloting an aircraft. Known neurological conditions may temporarily or permanently impair pilot cognition and function. For this reason, regulatory agencies may request neuropsychological testing as part of an assessment after a pilot experiences a neurological condition, illness, or injury or when questions arise regarding cognitive function that could impair their ability to fly an aircraft.

Although need for neuropsychological fitness in real world aircraft operations is universally accepted, methods of neuropsychological assessment can vary. Neuropsychological assessment of fitness to fly is a complex endeavor that includes testing, clinical assessment, and judgement. The ability of tests to accurately identify brain dysfunction is an important component of clinical assessments designed to reduce the probability of pilot error.

Purpose and Scope

The goal of this review is to identify, summarize and evaluate the available evidence on existing neuropsychological tests that have been used or proposed as part of assessments to determine whether pilots are fit to fly. Neuropsychological tests are tasks designed to measure the behavioral expression of brain dysfunction. Tests are used by a wide range of professionals to identify, assess, and treat brain impaired patients and may also be used to evaluate the effectiveness of treatment.[1] Based on consultation with stakeholders and researchers who have worked in this area, we limited the scope to neuropsychological testing of active pilots, and specifically focused on approaches to evaluation of function in experienced pilots when questions arise regarding neuropsychological function for any reason including illness, injury, medications, or substance abuse.

The intended purpose of this review is to inform both current policy discussions and future research. A comprehensive review of relevant evidence is an important first step in further developing and promoting ongoing rigorous research in this field and towards achieving the common goal of optimizing evaluation of neuropsychological function related to fitness to fly. Future efforts may include expanding the review to include evidence about neuropsychological testing of other populations (e.g., other transportation workers, other high risk professions, or pilot applicants and trainees). Also, it is likely more primary data collection and analysis will be needed to fully answer emerging questions about the best way to evaluate the neuropsychological function of experienced pilots who experience a neurological condition.

Key Questions

This review was designed to address the following questions about existing research on neuropsychological tests used in the field of pilot assessment. The objective is to help inform future policies and research about approaches to assuring the optimal pilot function and safe operations. :

1. What neuropsychological tests have been studied in active pilots to determine fitness to fly?
2. What is known about the psychometric properties of these tests?

- How well do tests predict pilot performance?
- How well do tests discriminate between pilots that are and are not impaired?

This review sought to identify and summarize research that directly answers these questions.

Methods

This review was conducted by the Pacific Northwest Evidence-based Practice Center (EPC) at Oregon Health & Science University. The EPC bases its work on standards established by the Agency for Healthcare Research and Quality and the Health and Medicine Division of the National Academies of Sciences, Engineering and Medicine (formerly known as the Institute of Medicine).[2,3] These methods are adapted to the topic and scope of each project. The specifics of the methods used for this review are described in this section.

Protocol Development and Registration

The key questions and the protocol, which details the methods, were developed by the research team with input from an Advisory Panel convened for this report. The panel consisted of 11 advisors including experts in the fields of neuropsychology, neurology, and methodologists in systematic reviews, and representatives of the Allied Pilots Association and Federal Aviation Administration (FAA) (See Appendix A for names and affiliations). Panel members disclosed financial and other conflicts of interest prior to participation and they contributed their expertise but did not officially represent the policies or positions of their employers.

The protocol was registered in the PROSPERO international database of prospectively registered systematic reviews (registration number CRD42018113045).

PICOTS and Inclusion and Exclusion of Studies

PICOTS is an acronym for Population, Interventions, Comparator, Outcomes, Timing and Setting. It is a framework that specifies the important elements that define the scope of the review. These are used to create the search strategy and become the basis for the criteria used to include or exclude a study from the systematic review. PICOTS operationalizes the elements of the key questions the review aims to answer. PICOTS does not necessarily describe the evidence that will be identified or what has been studied. Often, important questions have not yet been studied and one role of a systematic review is to highlight when this is the case.

The PICOTS for this review were:

- **Population:**
  - Pilots
    - Any age
    - Any type: commercial including air carrier/transport, military, or general aviation
- **Intervention** (tests):
  - Any neuropsychological test that has been used, or proposed for use, as part of an assessment of brain function in pilots.

Tests are limited to those that are focused on neurolopsychologic function and cognition. Assessments of other factors (e.g. fatigue, personality) were not included. These are listed in the exclusion criteria in Appendix B.

- **Comparator** (approach to evaluation of tests):
  - Comparisons to gold standard (e.g., diagnosis of cognitive dysfunction, functional performance)
  - Comparisons across tests
  - Comparisons across populations (e.g., pilots of different ages, experience)
  - Comparisons to generally accepted cut-off or threshold values or norms
- **Outcomes** (the results/values of the evaluations of tests):
  - Sensitivity, specificity, area under the receiver operating curve (AUROC)
  - Measures of predictive utility and discrimination
  - Contributions to explaining variance (e.g., output of regression models)
  - Correlations
- **Timing:**
  - Testing for screening or assessment following a known neurological condition, illness, injury or event, or when questions arise regarding neuropsychological function
- **Setting:**
  - Outpatient, occupational screening and/or evaluation

In addition to PICOTS, we made the following additional decisions a priori about what research studies to include.

Study Designs: We included comparative or predictive studies of any design including prospective and retrospective cohort studies, pre-post assessments, and cross-sectional studies of different populations. We excluded descriptive studies that provided information about a test or assessment, but no data on how the test or assessment performed in pilots. We also excluded commentaries and letters.

Language of publication: We restricted inclusion to English-language articles and presentations.

Sources: We limited inclusion to studies or reports published in 1980 or later. After consultation with stakeholders it was decided that changes in aviation and medicine would make studies conducted prior to 1980 less relevant. We started with studies published in peer-reviewed journals, but we also included government or technical reports, white papers, theses, and presentations at meetings if we could obtain the report or presentation and it included relevant data. We included conference abstracts only if the abstract contained usable data and we were not able to obtain the full presentation or a publication reporting the study results and the same data.

Detailed inclusion and exclusion criteria are included in Appendix B.

Literature Search Strategy

An overview of elements of the literature search are outlined below. The search strategies for the citation databases were developed and conducted by a specialist librarian and are included in Appendix C.

Publication Date Range: We searched for studies published from 1980 through March 15, 2019.

Citation Databases: Ovid MEDLINE®, PsycINFO, and Scopus.

Hand Searching: Reference lists of included articles and selected excluded articles (e.g., systematic and narrative reviews and descriptions of tests) were reviewed to identify additional, potentially relevant studies.

Grey Literature: Sources for grey (unpublished) literature included reports produced by government agencies or other organizations. These are not often indexed or easily accessible. To identify these types of publications, team members searched clearinghouses and conducted internet searches of OpenGrey, Google Scholar, FAA Aerospace Medicine Technical Reports, and the National Technical Reports Library. The strategies for these searches are also included in Appendix C.

Study Selection

The PICOTS framework described above was used to determine eligibility for inclusion and exclusion of abstracts. To ensure accuracy, all titles and abstracts were independently reviewed by two members of the research team who were blinded to each other's initial decisions. We retrieved the full text of articles deemed potentially appropriate for inclusion by at least one of the reviewers based on the abstract. We also reviewed the full text of any articles suggested by members of the Advisory Panel, reviewers of the draft report, or any other experts we consulted. Each full-text article was independently reviewed for eligibility by two research team members. Any disagreements about inclusion or exclusion were resolved by discussion with the entire research team until consensus was reached. A list of the included studies is provided in Appendix D. The studies that were excluded after the full text was reviewed are listed in Appendix E along with one reason for exclusion. Studies often met several exclusion criteria; however, as one is sufficient we did not list all possible reasons for exclusion.

Data Management and Extraction

Our inclusion and exclusion decisions were tracked using files created in EndNote and Microsoft Excel.

We also extracted key information from publications in Excel files in order to link publications reporting on the same study population, and organize results. Key data extracted from publications include: first author, year of publication, geographic location, type of pilots studied, sample size, demographic information about the pilots or study subjects (e.g., age) if reported, pilot flight experience, details on the test(s) evaluated, and the results of these evaluations. Sources of funding for studies were also recorded when reported. All study data were extracted by one research team member and verified for accuracy and completeness by a second research

team member. The data extraction tables are available as supplemental material from the report authors.

Appraisal of Individual Studies

A core requirement of a high quality systematic review is that in addition to identifying evidence, it must evaluate the evidence. This evaluation includes the appraisal of individual studies, which usually focuses on internal validity and potential sources of bias. However, the exact criteria used must be appropriate for the study design and the purpose of the review. Tools and criteria lists have been developed and tested for a range of designs (e.g., trials, observational studies, studies of diagnostic tests).

Our preliminary review of the literature revealed that a wide range of approaches and study designs have been used in research on neuropsychological testing of pilots. Unfortunately, a single previously validated and reliable tool was not available that could be applied across these diverse studies. In order to avoid using different instruments and criteria that would make it hard to summarize our appraisal of the included studies, we selected a small number of criteria that could be applied across studies with different designs. We reviewed available tools and approaches and selected four criteria we considered relevant to understanding the available evidence and the need for future research.

The four appraisal criteria used and reported in this review are:

1. Was the sample size greater than 30?
   Larger studies have more power to detect smaller differences and provide more precise estimates, though very large effects can be detected with a small number of subjects. We established this criterial as studies with 30 or fewer subjects may require use of different analytic approaches designed for small samples. Additionally, small studies have more potential for selection bias, may have limited generalizability, and can be difficult to replicate.
2. Was the study comparative or predictive?
   Studies that either compare tests or determine if a test predicts specified outcomes provide more relevant, direct information about a test. Information about the performance of a single test and associations (e.g. correlations) of test scores and outcomes provide less directly relevant information for the review questions. Using these requires judgements on whether the studies are comparable and the subjective interpretation of results.
3. Did the study address potential confounding in any way?
   Most of the included studies are observational and the results may be subject to different types of bias. Analyses should be designed to address potential sources of bias. This can be accomplished by stratifying results by potentially confounding variables and comparing results or adding variables in analysis models to control for differences. Studies that address confounding may still be affected by unidentified bias, but at least they acknowledge and address some sources of bias.
4. Did the study categorize subjects (e.g., as passing or failing, impaired or not, normal or abnormal etc.)?

This criteria addresses whether a study reported a test's ability to discriminate or categorize the subjects (pilots) into groups. Studies that use continuous outcome variables (e.g., a score on a simulation exercise) but do not establish cut-offs (e.g, what is pass or normal vs. what is fail or abnormal) do not directly address the review questions making them less relevant and their results difficult to interpret in the context of our questions.

These criteria were used to describe the studies in terms of selected elements of quality and their ability to directly address the key questions asked by this review. These criteria were not used to exclude studies; rather the criteria were used to describe included studies and identify a subset of the most informative studies for our designated questions.

The process for study appraisal was similar to the process for inclusion decisions. Two members of the research team assessed each study independently, compared responses, and reconciled differences. If the two raters were unable to reach consensus, the assessment was reviewed by the research team as a group and the assessment endorsed by the majority was used and reported.

Data Synthesis

Based on the data extracted from each study, we identified and reported study characteristics and results grouped according to the neurocognitive domain measured by the test.

We recognize that many tests evaluate multiple areas of functioning and that there is currently no established classification structure in the field. However, in order to help organize the information across a large number of different tests, we categorized each test as belonging to one domain for this review.

We selected six domains based on a leading textbook: attention; executive function; memory; motor performance; perception; reasoning.[1] In attention we included tests of concentration, mental tracking, and most reaction time tests. Memory is the domain we assigned to the retention of information. Many tests described in the studies as short-term memory tests have been categorized into the attention domain. Attention and memory are distinguished in that immediate recall is related to attentional capacity while memory is tested when there is a delay or interruption between acquiring and recalling information. In executive function we included tests which required planning or learning changing rules. Perception included tests that require mental manipulation of objects or recognition of abstract designs or figures. We included in the reasoning domain tests involving mental flexibility and concept formation. As situational awareness (SA) requires functioning across many domains and assessments of SA differ from those in traditional neuropsychiatric tests, we added this as a separate, seventh domain.[1]

Each test was assigned a domain according to its primary focus based on our interpretation of the test description provided in the included studies. We acknowledge that this is subjective and requires reducing complex tests to one domain. Appendix G presents the tests arranged by the domain we assigned them to in order to allow others to review and potentially revise the classification for future research.

As our preliminary review of the available evidence revealed that studies of this topic use variables approaches to evaluation of the tests, we did not expect to be able to combine results

and use quantitative synthesis (i.e., meta-analysis). Instead we providee summaries that include the ranges of results across groups of studies and strive to clearly describe our conclusions and their basis.

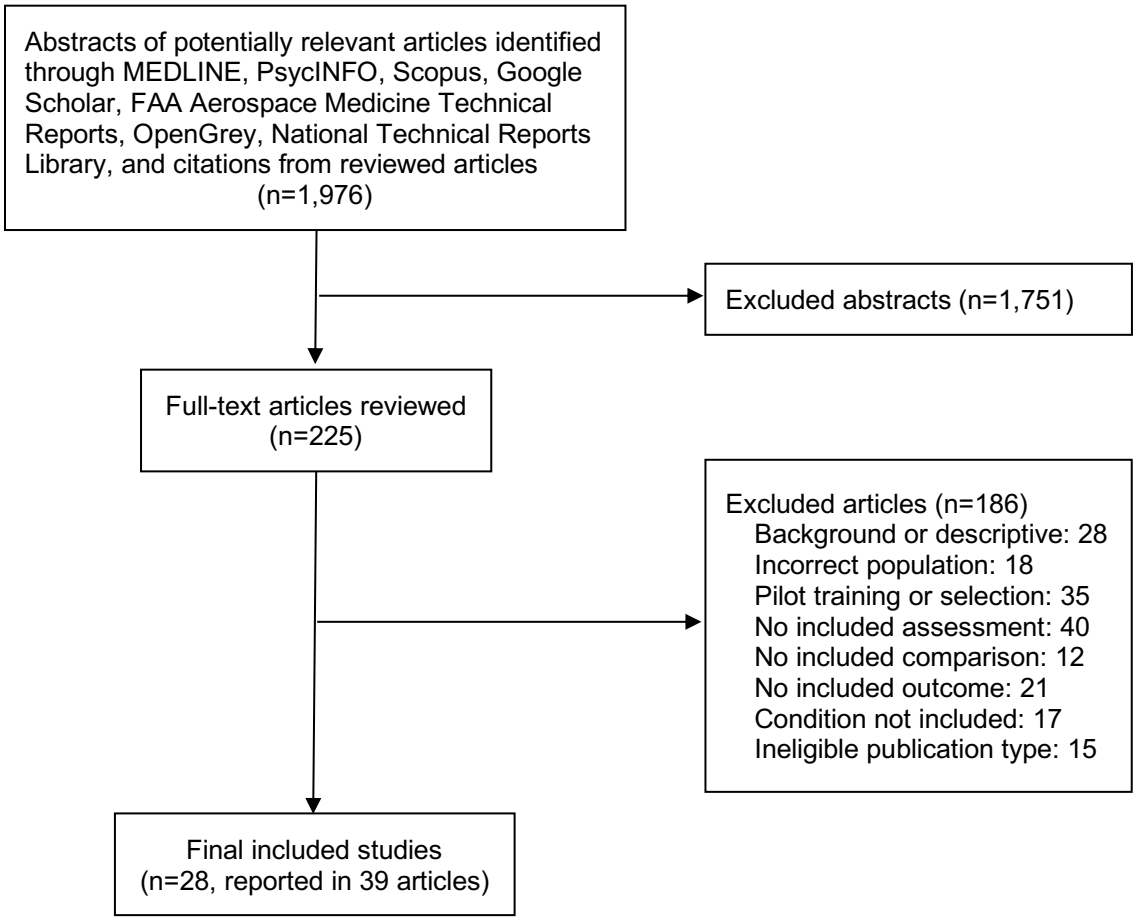
Results

Literature Search Yield

The results of the literature search, triage of abstracts, and review of full-text articles is summarized in the study flow diagram (Figure 1). Our searches yielded 1,976 potentially relevant citations. After reviewing titles and abstracts, 1,756 were excluded and the full text of 225 articles were pulled for review; of these, 39 articles that reported the results of 28 studies met our inclusion criteria. A list of the included studies is provided in Appendix D.

The excluded studies and the primary reasons for exclusion are in Appendix E. Many studies were excluded because they provided only descriptive information and did not evaluate the performance of the test. Other frequent reasons for exclusion were that the tests were not evaluated in pilots or that the data provided or the analysis did not permit conclusions about test performance.

Figure 1. Literature flow diagram



Description of Included Studies

Key characteristics of the included studies are represented in the charts in Figures 2-5. For more than half of the included studies, the first article was published in the 1990s. This likely reflects the attention to related topics such as age limits for pilots at the time and the publication of research funded by the FAA, including the work leading up to the development of the CogScreen: Aeromedical Edition. Almost three quarters of the studies were conducted in the US studying US pilots. Military pilots were the subject of 36% of the studies, though 21% of the studies did not provide information on the type of pilot. While 64% of the studies analyzed samples of 51 or more people, 36% would be considered small with sample sizes of 50 or fewer subjects. When studies included multiple analyses containing data on different numbers of people we classified the study based on the largest sample size.

The data used to produce these charts are included in Appendix F. This Appendix includes two tables. The first, Table F-1 provided the counts and percentages of these characteristics across the studies. The second, Table F-2 provided the information for each included study.

Figure 2. Types of Pilots Studied

(percentage of included studies)

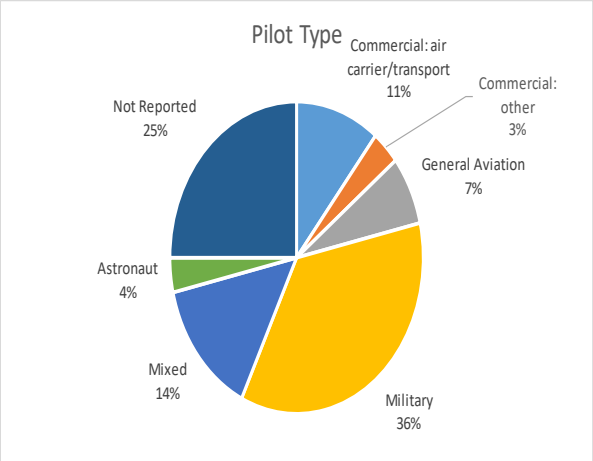


Figure 3. Decade of Publication

(number of included studies)

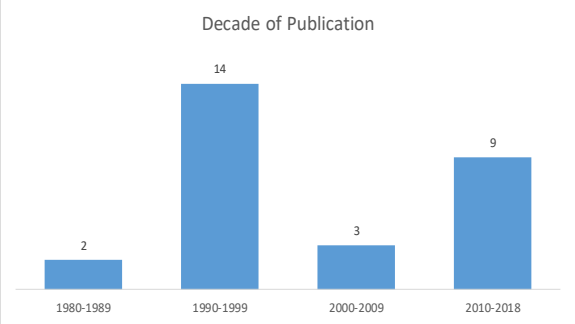


Figure 4. Geographic Location

(percentage of included studies)

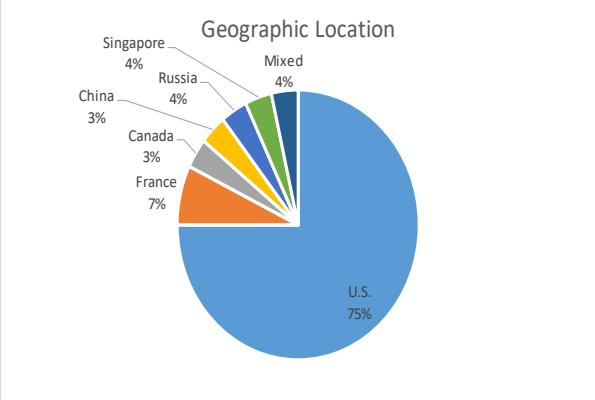
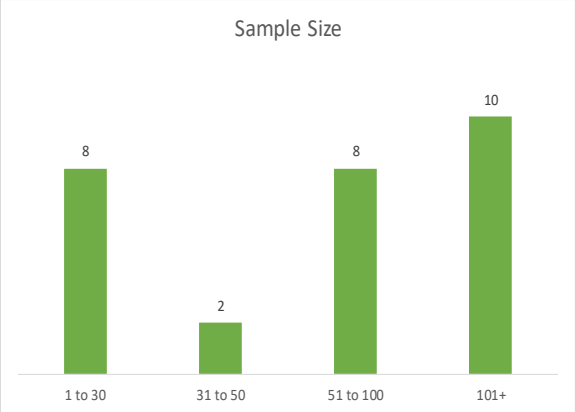


Figure 5. Sample Size

(number of included studies)



Neuropsychological Tests Evaluated

Most of the included studies evaluated several different neuropsychological tests. Some studies evaluated discrete tests, designed to assess one neuropsychological function while others evaluated test batteries containing multiple tests or selected tests from a battery. A complete list of the 106 tests evaluated in the included studies and the batteries which they are part of (if applicable) is provided in Appendix G.

We grouped the tests into seven domains (as described in Methods above) and three categories when multiple tests were evaluated together: formal test batteries, subsets of batteries, and combinations of tests (not formal batteries). When individual tests that sought to assess multiple domains were formally combined we labelled these batteries (groups of tests) and we also noted where subsets of batteries were studied rather than the entire battery. In some instances, studies combined tests that are not part of any formally constructed test batteries, and which belonged to different domains. These we classified as "combination". Table 1 lists the neuropsychological domains and these categories, reports the number of tests evaluated, and provided corresponding citations. The most frequently evaluated test domain in the included studies was attention with some assessment of attention included in 61% of the studies (17 studies). Reasoning was the next most frequently evaluated domain and was included in 46% of the studies (13 studies). Batteries or subsets of batteries were assessed in 15 (54%) of the included studies. The least frequently studied domain was situational awareness.

Table 1: Tests Evaluated in Included Publications by Neuropsychological Domain

Neuropsychological Domain	Number of Tests*	Citations
Attention	34	4-30
Reasoning	17	4-7,9,10,12,13,15-18,20-28,31,32
Perception	15	9,10,12,13,21-26,30-32
Batteries	12	9,12,13,15-18,33-37
Memory	7	8-10,15-18,20,32
Combination	7	4-6,12,13,21-26,30,38
Motor performance	5	4-7,11,15,16
Battery subset	3	12-14,17-19,21-28,31,37,39
Executive function	3	9,10,13,14,19,21-28,32
Situational awareness	3	9,10,20,40-42

*publications reported evaluations of multiple tests and each test is counted separately in this table.

Study Design: Outcomes and Analytic Approaches

Tables 2 and 3 provide information about key elements of the study design, outcomes and the analytic approaches used in the included studies.

Outcomes

In studies that evaluate a test, the researchers have to decide what to use as the "outcome." In this situation the outcome is what the test result is associated with or what the test result predicts. The most frequently used outcome was pilot performance in a simulator. Simulatior performance was the outcome in 17 (44%) of the included research reports. Additional outcomes included indicators of situational awareness (5 studies), the presence or absence of a diagnosis (5 studies), or indicator of a brain injury (3 studies). Five studies used pilots' actual performance while flying a plane based on data obtained through analysis of flight recorders or evaluations and observations of supervisors. The citations indicating which publications included each type of outcome are provided in Table 2.

Table 2: Outcomes

Outcome	Publications #	Publications %	Citations
Simulator Performance (e.g., flight path deviations, maneuvers)	17	44%	5-7,9,20-28,34,39,40,42
Clinical status: condition or diagnosis	5	13%	12,13,16,29,36
Situational Awareness	5	13%	4,8,10,20,32
Performance (black box, job performance)	5	13%	11,13,14,30,31,38
Other Outcomes	5	13%	13,15,33,35,37,41
Brain injury indicators	3	8%	17-19

Notes: 1. Kay, 1995: CogScreen manual is in the publication count for "clinical status or diagnosis" only, but it also provides additional data on studies reported in Kay 1991, Kay 1993, and Yakimovich 1994. 2. Publications may have performed analyses for more than one outcome. 3. The percentage is out of 39 included publications.

Analytic Approaches

There are several analytic approaches that can be used to assess a test including measuring its association with an outcome or modeling the extent to which a test predicts the outcome. Different analytic approaches provide different types of information and allow (or don't allow) for different conclusions. For example, simple correlations cannot convey if a test can discriminate or "sort pilots" into groups, they only measure the extent that changes in a test score are reflected in changes in an outcome. Complex analytic models can provide more information, but their ability to directly address our questions still depends on the details. Twelve studies used regression models for their analyses. The advantage of these models is that they allow other variables to be added (or controlled for) in the analysis. For example, a model may include age and type of pilot certification as well as the test score in predicting simulator performance. However, linear regression models that use a continuous outcome as the dependent variable, still provide only information about the relationships among the variables and do not provide

information on the ability of a test to sort pilots into groups. In contrast, logisitic regressions and other approaches that assess diagnostic accuracy (whether a test can provide a correct diagnosis) and discriminant ability directly address our questions about test performance for pilot evaluation.

Table 3: Analytic Approaches

Analytic Method	Publications #	Publications %	Citations
Correlations	12	31%	4,8,10,11,17,19,20,31,34,38,40,42
Regression models	7	18%	5,13,14,21-28,30,32,39,41
Discriminant analysis / Diagnostic accuracy	5	13%	6,12,13,16,37,38
Other	5	13%	9,15,18,29,33
ANOVA/ANCOVA	3	8%	7,13,35

ANOVA = analysis of variance; ANCOVA = analysis of covariance
Notes: 1. Linear mixed effects (longitudinal) were classified as regression models; Other includes chi-squared, t tests, and descriptive statistics. 2. The percentage is out of 39 included publications.

Appraisal of Individual Studies

We used four criteria (described in the Methods section above) to help us identify the most rigorous studies and create a descriptive profile of the evidence. Table 4 reports the results of that process in two ways. First, the number of articles that met each individual criterion are reported on the left side of the table in the first two columns. This reports how common meeting each criteria is in the included studies. Most of the included articles reported on studies that were predictive or comparative (31 of 39) but only about 1/3 (13 out of 39) categorized subjects in some way. Second, the number of articles that met each possible number of criteria - all four criteria, three, two, one or none are reported on the right side of the table. This allows us to identify the strongest evidence, specifically the subset of studies that met all four criteria across all the included studies. Our appraisal of each study is included in Appendix I.

Table 4: Summary of Appraisal of Study Design Elements

Criteria	# of articles meeting criteria		Total # of Criteria Met	# of articles with this profile
Sample size >30	29		Four	8
Predictive or comparative design	31		Three	14
			Two	8
Confounding addressed	23		One	6
Categorized subjects	13		None	3

Findings from Included Studies

In order to provide different perspectives and summarize the findings across the identified studies, we organized and present the results in three ways: first, we provide information from the eight studies that met all four of our appraisal criteria; second, we provide the results when individual tests are compared head-to-head; and third, we present the primary findings from all included studies, organized by neuropsychological test domain.

Studies Meeting all Four Appraisal Criteria

Tables 5 and 6 provide an overview of the eight studies that met all four of our appraisal criteria. These studies represent those providing the most rigorous data for addressing our key questions and assessment. It is important to realize that studies may provide stronger or weaker evidence depending on the question being asked. For example, studies that examined the relationship between age and neuropsychological tests and provided data were included, but they often did not meet the criteria of categorizing subjects in a way related to fitness to fly. These studies may have been well designed and conducted to answer questions about age, but they may not be as useful for this review given our focus on evaluating the ability of tests to identify impairments or predict pilot performance.

Table 5 provides an overview of the neuropsychological domains assessed in these eight studies. All these studies assessed multiple tests that included different domains or a battery comprised of multiple tests.

Table 5: Neuropsychological test domains included in studies meeting all appraisal criteria

	Attention	Executive Function	Memory	Motor	Perception	Reasoning	Combination	Battery	Battery subset
Causse, 2011b	X			X		X			
Kay, 1991, 1995 Phase B	X	X				X		X	X
Kay, 1995 Phase C Clinical	X				X	X	X	X	X
O'Donnell, 1992	X		X	X		X		X	
Stokes, 1991								X	
Taylor, 2000									X
Tolton, 2014	X	X				X			X
Yakimovich, 1994; Kay, 1995	X	X							X

Table 6 provides a brief summary of the key finding(s) for each of these studies. This table demonstrates that the research to date is highly variable in that different domains and tests have been evaluated, and the approaches to evaluation have varied.

Table 6: Findings from articles meeting all four selected appraisal criteria

Author, Year Outcome	Summary of Key Findings
Causse, 2011b[7] Simulator Performance	*Specific Outcome: Crosswind landing decision* Attention: 2-back test, Significant Reasoning Wisconsin card sorting test, Significant *Specific Outcome: Flight path deviations* Attention Spatial Stroop test, NS Motor Performance: Target hitting test: Significant Reasoning Wisconsin card sorting test: NS Reasoning test: Significant
Kay, 1995[13] "Phase C Clinical" Clinical status or diagnosis	CogScreen battery subsets *Discriminant function model (using sample 1)* 53% accuracy overall *Logistic Regression Probability Value (LRPV) model (using sample 2)* Diagnostic accuracy: Sensitivity: 83%. Specificity: 95%. Accuracy: 90% Estimated probability of brain dysfunction: 81% in clinical group with a diagnosis; 12% in pilot group; 27% of normative sample
Kay, 1991[12], 1995 Phase B[13] Clinical status or diagnosis	CogScreen battery *Diagnostic accuracy* (≥1 test score at <5th percentile) Sensitivity 73% Specificity: 90% PPV: 0.78 with prevalence of 49.4% CogScreen battery subset *Discriminant function model using speed and accuracy measures* (≥1 test score <5th percentile) Accuracy: 89% overall: 90% in pilots; 88% in patients Reasoning Patients scored lower on both tests of reasoning compared to pilots or non-pilot subjects (ANCOVA analysis p≤0.01): Math and Pathfinder
O'Donnell, 1992[16] Clinical status or diagnosis	Development study (2 versions of a test being developed) Neuropsychological Test Battery Discrimination and Classification Accuracy Version 1: 68% of variables discriminate clinical status Sensitivity: 95.00% (95% CI: 83.08% to 99.39%) Specificity: 92.59% (95% CI: 84.57% to 97.23%) Accuracy: 93.39% (95% CI: 87.39% to 97.10%) Calculated from: 38 true positives, 6 false positives, 2 false negatives, 75 true negatives. Version 2: 59% of variables discriminate clinical status Classification not reported
Stokes, 1991[36] Clinical status or diagnosis	SPARTANS battery is best discriminator among tested neuropsychological batteries SPARTANS Sensitivity: 66.3% Specificity: 87.3% PPV: 0.867 Accuracy: 75.9% AMA mini-mental test Sensitivity: 64.5% Specificity: 81.1% PPV: 0.799 Accuracy: 72.2% Illinois Screening Test, version 2 Sensitivity: 72.8% Specificity: 74.2% PPV: 0.766 Accuracy 73.3%

Author, Year Outcome	Summary of Key Findings
Taylor, 2000[43] Simulator Performance	CogScreen-AE subset 45% of Simulator Performance variance explained by 4 CogScreen factors. F(4, 77)=16, p<0.0001
Tolton, 2014[27] Simulator Performance	CogScreen subsets *Low workload vs. High workload* Working Memory measures: 32.7% vs. 24.4% contribution to variance in performance Processing Speed measures: NS for either Tracking measures: p=0.10 and NS CogScreen-AE LRPV: NS in model with pilot factors; 51.9% scored above level suggestive of brain dysfunction.
Yakimovich, 1994[14] Kay, 1995[13] Performance	CogScreen subset *Multiple regression models:* Results varied by aircraft TU-154 predicted by 3 tests: Divided Attention, Dual Tracking, and Shifting Attention, 30% of variation explained by model (adjusted multiple R^2=0.30) IL-86 predicted by 3 tests: Pathfinder, Backward digit span, and Dual tracking 46% of variation explained by model (adjusted multiple R^2=0.46)

AE = Aeromedical Edition; AMA = American Medical Association; ANCOVA = analysis of covariance; CI = confidence interval; NS = not statistically significant; PPV = positive predictive value; vs. = versus

Head-to-Head Test Comparisons by Domain

The variation in the approaches to evaluating the tests in different studies makes it difficult to directly compare results and attempt to conclude whether certain tests or domains are more or less useful as part of an assessment for neuropsychological impairment. An approach to considering the comparative performance is to focus on studies that compare two or more tests for pilots with similar problems.

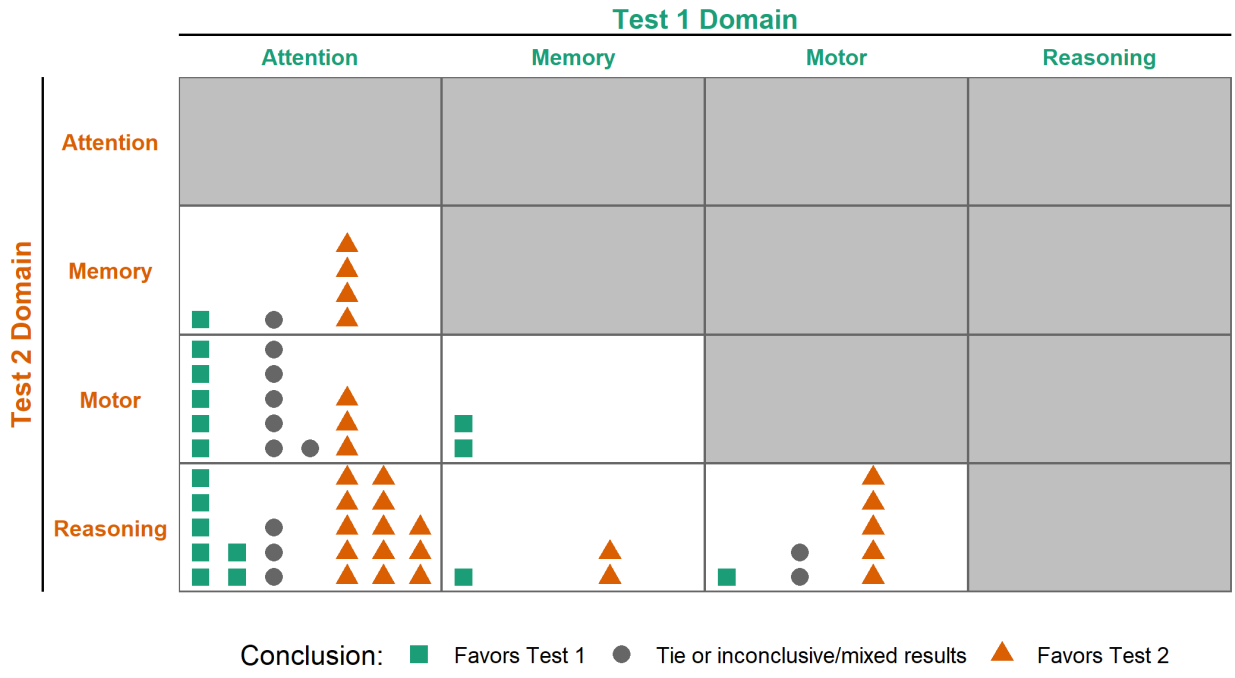
To assess comparative test performance, we first selected only the studies that categorized the subjects (our fourth appraisal criterion) so that the type of comparison is the same, but we allowed the analytic approach and outcomes to be different. We then identified those which provided results for tests in different domains; we excluded batteries, battery subsets, and combinations, as these are not limited to a single domain. We compared the test's ability to distinguish between groups for each pair of tests from different domains (e.g. a test of attention compared to a test of reasoning). Each comparison was coded as favoring one of the two tests or as tie/inconclusive if the tests' performance was not different.

A chart visualizing the relative performance of different domains enables identification of trends in results (Figure 3). Tests are grouped by domain (domains for Test 1 as columns, and domains for Test 2 as rows), and the conclusion from each comparison is represented by a symbol. In addition, the domain names and symbols are color-coded (green for Test 1 domains and

conclusions favoring Test 1, orange for Test 2 domain and conclusions favoring Test 2, and gray dots for tie or inconclusive/mixed results).

Figure 3. Comparative performance of tests in different domains for classification of subjects

(Note: cells on the diagonal and above are intentionally blank as they would repeat the results in the cells below).



Of the 13 studies that categorized subjects, there were three that provided data comparing tests from different domains, resulting in 56 comparisons.[6,7,16] None of the studies had data allowing for comparison of any tests in executive function, perception, or situational awareness. Among the four domains compared, the most frequent comparison was attention vs. reasoning (23 comparisons), followed by attention vs. motor (14 comparisons). While no definitive pattern emerges, there may be some trends. Tests of attention more often perform better than tests of memory and reasoning, but comparisons against motor do not show a clear advantage for either. Reasoning tests appear to outperform motor tests, though there are limited comparisons. However, our confidence in this finding is low as these data are from a very small number of studies.

This approach and Figure 3 is limited in that it provides information on the comparative performance of the test in the population for that study. However, it does not necessarily provide information that can be generalized to the comparative performance of the tests in different populations of pilots. This is particularly important to acknowledge because the underlying decifit that we want to assess in a pilot may be different. Specifically, a test may be useful for pilots referred for one condition such as a stroke but less useful if they were referred for a different condition such as an alcohol use disorder.

Results by Domain

In the next sections the results of the included studies are split into two groups: evaluations of multiple tests grouped together and evaluations of individual tests. If a study evaluated both a battery and individual tests, the results are split into the corresponding tables. The purpose of these sections is to group the results and allow them to be viewed together. Given the differences in design, analysis and outcomes, quantitative synthesis (meta-analysis) is not possible. The intention is that this grouping will allow the reader to understand our judgements about the trends or overall direction of the evidence or facilitate their own judgements.

The information in these tables is limited to outcomes, test and key results. Each table contains a row for each outcome-assessment pair and the corresponding results. More information on the studies, such as sample size and type of pilot are included in Appendix F.

Batteries, Subsets of Batteries and Combinations of Tests

Tables 7, 8 and 9 summarize the results of studies that evaluated batteries, subsets of batteries and studies that evaluated combinations of tests that were not constructed as batteries. As these all combine multiple tests, they may be, though are not required to be, more comprehensive. Test batteries vary considerably in length and complexity, ranging from a few minute to several hours to complete. Some batteries seek to combine measures that will assess a wide range of relevant cognitive skills, while others focus on a single or limited number of cognitive functions. Appendix G includes a list of the tests evaluated in the studies included in this review and indicates the battery in which it is included, if applicable. Appendix H is a glossary that provides a brief description of the tests and batteries evaluated in the included studies.

Table 7 provides an overview of the six studies that evaluated full test batteries. Studies that included only subsets of items from batteries are included in Table 8. These studies included assessment of the CogScreen, a battery created specifically to evaluate pilots, as well as other

studies of shorter mini-mental tests created for general screening. Most of these publications (Kay, O'Donnell and Stokes) report evaluations conducted as part of the development of what were new or proposed batteries at the time of the study. The studies by Kay (1995) were part of the development of the CogScreen, while O'Donnell (1992) developed The Neurosychological Test Battery. Both of these evaluated the set of tests in terms of their ability to discriminate, or sort people into healthy or impaired. Stokes (1991) took a similar approach but also evaluated two common mental status tests to allow comparisons of tests designed to be used for standard screen and diagnosis in the general population to ones designed for pilots.

Table 7: Batteries

Outcome	Assessment	Primary test performance results	Author, Year
Brain injury indicators	MicroCog	Correlations all: NS (neurologic decompression sickness, total hours or frequency of hypobaria)	McGuire, 2014[18]
Brain injury indicators	Multidimensional Aptitude Battery–II	Correlations all: NS (neurologic decompression sickness, total hours or frequency of hypobaria)	McGuire, 2014[18]
Group comparison or classification	CogScreen	Mixed sample of pilots and non-pilots, all clinical patients (alcohol, aviation performance and psychiatric referrals; confirmed neurologic disorders, suspected neurologic disorders) Scores different across clinical only groups **p<0.001** Neurologic patients (n=40) vs. age and education matched pilot (n=60) Mean estimated probability of brain dysfunction Clinical group: 0.81 (SD 0.18) Pilot group: 0.12 (SD 0.23) Entire aviator normative sample: 0.27 (significantly different by age groups - age <45 years: 0.18 (SD 0.26) - age ≥45 years: 0.37 (SD 0.34) Correct classification Normal pilots: 95%. Specificity Mild brain dysfunction: 82.9% Sensitivity Classification accuracy: 90.1%	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1)[13]
Group comparison or classification	CogScreen	Pilots and healthy non pilots vs. patients 17 of 22 CogScreen measures =significant Pilots vs. healthy non pilots: none significant Diagnostic accuracy: positive test = at least one CogScreen test score below 5th percentile - **Sensitivity 73% (true positive rate 29/40)** **- Specificity: 90% (true negative rate 37/41)** **- PPV: 0.78 (prevalence = 49.4%)** - positive tests in 82.5% of patients (33/40) and 34% of pilots (14/41)	Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study")[13]

Outcome	Assessment	Primary test performance results	Author, Year
Group comparison or classification	CogScreen	Pilots on selective serotonin reuptake inhibitors, approved vs. denied medical certification based on neuropsychological testing Diagnostic accuracy: positive = at least one CogScreen test score below 15th percentile prevalence (denied certificate) = 11.2% some results calculated for this report: - Sensitivity: 100.0% (95% CI 71.5 to 100.0) - Specificity: 87.4% (95% CI 78.5 to 93.5) - LR+: 7.91 (4.55 to 13.74) - LR-: 0.00 (95% CI NA) - Accuracy: 88.8% (95% CI 80.8 to 94.3) - PPV: 50.0% (95% CI 36.5 to 63.5) - NPV: 100.0% (95% CI NA)	DeVoll, 2013[37]
Group comparison or classification	Neuropsychological Test Battery	Test Development Discriminating non pathology group from group with known pathology Version 1.0 **42 of 62 test variables significantly discriminated p<0.05** Version 1.1 Identified 95% of True Positives, 5% of False Negatives, 7.4% of False Positives calculated for this report: Sensitivity: 95.00% (95% CI: 83.08% to 99.39%) Specificity: 92.59% (95% CI: 84.57% to 97.23%) Accuracy: 93.39% (95% CI: 87.39% to 97.10%) Version 2.0 **19 of 32 test variables significantly discriminated p<0.05**	O'Donnell, 1992[16]
Group comparison or classification	Illinois Screening Test version 1	Discriminate between pilots and non-pilots with several different clinical diagnoses Sensitivity: 72.6% Specificity: 83.3% PPV: 0.833 Accuracy: 77.59%	Stokes, 1991[44]
Group comparison or classification	Illinois Screening Test version 2	Discriminate between pilots and non-pilots with several different clinical diagnoses Sensitivity: 75.8% Specificity: 88.9% PPV: 0.887 Accuracy: 81.90%	Stokes, 1991[36]
Group comparison or classification	FFM mini-mental test	Discriminate between pilots and non-pilots with several different clinical diagnoses Sensitivity: 54.8% Specificity: 77.7% PPV: 0.850 Accuracy: 70.69%	Stokes, 1991[44]
Group comparison or classification	AMA mini-mental test	Discriminate between pilots and non-pilots with several different clinical diagnoses Sensitivity: 69.3% Specificity: 77.7% PPV: 0.787 Accuracy: 74.14%	Stokes, 1991[44]

Outcome	Assessment	Primary test performance results	Author, Year
Group comparison or classification	SPARTANS	Discriminate between pilots and non-pilots with several different clinical diagnoses **Sensitivity: 82.3%** **Specificity: 96.3%** **PPV: 0.962** **Accuracy: 88.79%**	Stokes, 1991[44]
Simulator Performance	CogScreen	Correlation: Total score with subjective evaluation of emergency abnormal maneuvers **r = 0.48, p<0.05** All other correlations: NS Including all simulator maneuvers, routine maneuvers, and challenging maneuvers with recorded deviations or subjective evaluation and emergency/abnormal maneuvers with recorded deviations.	Hyland, 1994[34]
Simulator Performance	Flitescript (situational knowledge in long term memory)	Correlations: NS Including all simulator maneuvers, routine maneuvers, challenging maneuvers, and emergency/abnormal maneuvers with recorded deviations or subjective evaluation	Hyland, 1994[34]
Simulator Performance	WOMBAT (ability to perform several tasks and change priorities)	Correlations: NS Including all simulator maneuvers, routine maneuvers, challenging maneuvers, and emergency/abnormal maneuvers with recorded deviations or subjective evaluation	Hyland, 1994[34]

CI = confidence interval; n = number; NS = not statistically significant; PPV= positive predictive value; SD = standard deviation
Bold = statistically significant

Table 8 includes the primary results from eight studies that evaluated subsets of test batteries. Some of these results are from studies that also evaluated the entire battery (Kay, 1995; McGuire, 2014). In these cases the goal was often to determine if a subset was driving the results when they are significant, or if a subset performed well but is 'hidden' in overall scores when the analysis provided inconclusive or insignificant findings. These subset analyses used the same outcomes; brain injury indicators and accuracy of classification, as they did for the analysis of the complete battery. Other studies used subsets either to reduce burden on study subjects (it may be feasible to do a 1 hour assessment in more people than a 6 hour assessment) or because the subset corresponds to the underlying construct they wanted to test. Most of these studies assessed the battery subsets in terms of their correlations with actual flight evaluations (Barron, 2016) or simulator flight performance (Kennedy, 2013; Taylor, 2000; or Tolton, 2014).

Table 8: Battery subsets

Outcome	Assessment	Primary test performance results	Author, Year
Brain injury indicators (imaging)	MicroCog: subset	Pilots with repeated hypobaric exposure vs control pilots General cognitive functioning domain**: p=0.04** General cognitive proficiency domain: NS Information processing accuracy domain**: p=0.032** Information processing speed domain NS Spatial processing domain: NS White matter hyper intensity (high vs. low count) General cognitive functioning domain**: p=0.020** General cognitive proficiency domain: **p=0.022** Information processing accuracy domain**:** NS Information processing speed domain NS Spatial processing domain: NS	McGuire, 2014[18]
Brain injury indicators	CogScreen: subset	CogScreen LRPV model Correlations duration of post-traumatic amnesia: **r = 0.56, p=0.01** loss of consciousness: **r = 0.50, p=0.01** initial GCS score: NS initial estimated seizure risk: NS	Moore, 1995[19]
Group comparison or classification	CogScreen: subset	Classification accuracy Accuracy measures (Math and Matching to Sample) Overall: 69% Subgroups Pilots: 70.7% (29/41). Patients: 67.5% (27/40) Response Speed (4 tasks) Overall: 76.5% Subgroups Pilots: 92.7% (38/41). Patients: 60% (24/40) CogScreen-AE discriminant function model (7 tasks) Overall 88.9% [72/81] Subgroups Overall excluding n=10 alcoholic patients: 94.4% Pilots: 90.2% (37/41) Patients: 87.5% (35/40)	Kay, 1995[13] (Kay 1991/Clinical Studies I "FAA Phase B Study")[12]
Group comparison or classification	CogScreen: subset	CogScreen LRPV model Pilots on selective serotonin reuptake inhibitors, approved vs. denied medical certification based on neuropsychological testing Diagnostic accuracy: positive = LRPV ≥0.8 prevalence (denied certificate) = 11.2% some results calculated for this report: - Sensitivity: 36.4% (95% CI 10.9 to 69.2) - Specificity: 87.4% (95% CI 78.5 to 93.5) - LR+: 2.88 (1.10 to 7.49) - LR-: 0.73 (95% CI 0.46 to 1.15) - PPV: 26.7% (95% CI 12.3 to 48.6) - NPV: 91.6% (95% CI 87.3 to 94.5) - Accuracy: 81.6% (95% CI 72.5 to 88.7)	DeVoll, 2013[37]
Performance	Air Force Officer Qualifying Test: subset	Correlation with Officer Performance Reports (OPR) Manned aircraft pilots 1st OPR corrected r = 0.175 **p≤0.01;** 3 OPR scores: corrected r = 0.168 **p≤0.01**; Remote pilots 1st OPR corrected r = 0.332 **p≤0.01;** 3 OPR scores: corrected NS	Barron, 2016[31]

Outcome	Assessment	Primary test performance results	Author, Year
Simulator Performance	CogScreen: subset	Speed of processing standardized composite Generalized linear models Dependent variable: approach aileron movements Effect size -0.26, beta coefficient -0.28 (SE 0.13), **p=0.036** Dependent variable: landing decision accuracy Effect size -0.26, beta coefficient -0.150 (SE 0.685), **p=0.029** Correlations with landing decision accuracy Older pilots: r = -0.27, p=0.11 Younger pilots: r = -0.17, p=0.32	Kennedy, 2010[39]
Simulator Performance	CogScreen: subset	Correlation Complex intra-individual variability: NS -Includes 5 tasks Linear mixed effect models Dependent Variable: Initial simulator performance Intra-individual variability: p**<0.0001** Processing Speed: **p<0.0001** Executive Function: NS Expertise: **p<0.0001** Dependent Variable: decline in simulator performance with age Intra-individual variability: NS Processing Speed: NS Executive Function: NS Expertise: NS	Kennedy, 2013[21]
Simulator Performance	CogScreen: subset	Processing speed (composite measure of 11 tasks of visual scanning and perceptual comparison) Correlations - overall summary score: r = 0.40, **p<0.001** - communication: r =0.42, **p<0.001** - traffic avoidance: r = 0.13, **p<0.05** - approach: r = 0.24, **p<0.001** - emergency: r = 0.218 **p<0.001**	Kennedy, 2015[22]
Simulator Performance	CogScreen: subset	CogScreen-AE (4 factors) - General Speed/Working Memory - Visual Associative Memory - Tracking - Motor Coordination overall flight simulator performance**: R^2 = 0.45, p<0.0001** Model: 4 factors combined	Taylor, 2000[43]
Simulator Performance	CogScreen: subset	<u>CogScreen LRPV</u> Correlation: r = -0.346, p=0.01 Regression model: NS when added to model with age, experience, processing speed and tracking	Tolton, 2014[27]

Outcome	Assessment	Primary test performance results	Author, Year
Simulator Performance	CogScreen: subset	General speed and working memory (average of 5 factors) Correlation: r = 0.279, **p=0.045** Regression models contained: working memory + processing speed + tracking + experience + age: High workload: only significant predictor: experience **p=0.041** Model**:** cumulative R^2 = 0.350, p=0.016 Low workload **Significant predictors: working memory, visual tracking, and expertise**. cumulative R^2 = 0.496, p<0.0001	Tolton, 2014[27]
Simulator Performance	CogScreen: subset	Speed and working memory composite Correlation**: r = -0.456, p<0.01** Multiple regression model standardized beta = -0.337**, p=0.004** adjusted for expertise, visual attention, cognitive flexibility Combination of 3 composites Speed and working memory, Shifting attention, Visual attention added to model with expertise only: change in R^2 = 0.410, **p<0.01** Overall model: speed/working memory + cognitive flexibility + visual attention + age + expertise: R^2 = 0.591, p<0.01	Van Benthem, 2016[28]

AE = Aeromedical Edition; GCS = Glasgow Coma Scale; LRPV = logistic regression probability value; NS = not statistically significant
Bold = statistically significant

Table 9 contains the primary results from studies that combined different tests, assessing them as a group, but the combinations were not preexisting batteries at the time of the studies. Again, some were part of exploratory efforts to identify domains for the development of batteries. However, most were studies designed to compare the relative contribution of neuropsychological functioning in certain domains and expertise to pilot performance. In one case (Shull, 1990) the outcome was flight performance while in most it was simulator performance. These studies report simple correlations between the combination or composites and the performance score and then create regression models designed to estimate how much variation in performance can be explained by these variables.

Table 9: Combinations - not formal batteries

Outcome	Assessment	Primary test performance results	Author, Year
Group comparison or classification	PASAT + Trail Making Test + Symbol Digit Modalities Test	Accuracy Overall: 80.25% (using pilots and patients only; 65/81) Pilots: 78.0% (32/41) Non-pilot healthy: 64.3% (27/42) Patients: 82.5% (33/40)	Kay, 1995[13](Kay 1991/Clinical Studies I "FAA Phase B Study")[12]
Performance	Combination: psychomotor and	Correlation analysis: NS	Shull, 1990[38]

Outcome	Assessment	Primary test performance results	Author, Year
	dichotic listening task measures	No significant relationship between test measures and performance during combat training exercises	
Simulator Performance	Model: 2-back test + deductive reasoning test + total flight experience	Regression Adjusted R^2 = 0.4451, **p<0.05** Model: working memory + reasoning + total flight experience	Causse, 2010[5]
Simulator Performance	Speed of processing composite	Correlation: r = 0.33, **p<0.05** General linear model difference between groups by expertise level: **p<0.02**	Taylor, 2005[25]
Simulator Performance	Speed of processing composite	General linear model difference by expertise level: p=NS	Taylor, 2007[24]
Simulator Performance	Speed of processing composite	Mixed effects growth curve analysis Initial flight simulator performance - processing speed main effects: **p=0.0006** - Processing speed x Executive function interaction: NS Rate of decline in flight simulator performance: - processing speed: **p=0.010** - Processing speed x Executive function: **p=0.008** Exploratory ROC analysis - Processing speed: kappa = 0.25, **p<0.001** - Processing speed + executive function: kappa = 0.19, **p<0.05**	Yesavage, 2011[26]
Performance	Cognitive ability tests	Hierarchical regression models Model: cognitive ability tests (tests entered together as first block of variables) + age + expertise Dependent variable: Aircraft route probes accuracy - **R^2 = 0.185, p<0.001** Dependent variable: ATC message readback **- R^2 = 0.226, p<0.001** Dependent variable: Route recall accuracy **- R^2 = 0.240, p<0.001**	Morrow, 2003[30]
Situational Awareness	General cognitive ability (GCA) composite	Partial correlation: r = NR**, p<0.05** adjusted for job experience measures (F-15 and total flying hours) Linear regression models Incremental R = 0.023, **p=0.01** Models compared: GCA composite + experience vs. experience alone Individual subtest: NS	Carretta, 1996[4]
Situational Awareness	Psychomotor (PM) composite	Partial correlation: r = NR**, p<0.05** adjusted for job experience measures (F-15 and total flying hours) Linear regression: NS Models compared: PM composite + experience vs. experience alone Individual subtest: NS	Carretta, 1996[4]

ATC = air traffic control; NR = not reported; NS = not statistically significant; PASAT = Paced Auditory Serial Addition Test; ROC = receiver operating characteristic; vs. = versus
Bold = statistically significant

Single Domains

Tables 10 – 17 each contain the results from evaluations of single tests reported in the included studies. These are sorted by domain, with a separate table for each domain: memory (Table 10); motor response (Table 11); perception (Table 12); situational awareness (Table 13); reasoning (Table 14); executive function (Table 15); attention, simple (Table 16); and attention, complex (Table 17). These tables contain more results than the head-to-head comparisons summarized in Figure 3, because many studies did not directly compare tests; they simply reported the correlations of a test with the outcome, whether the test was a significant predictor in a multivariable model, or the accuracy with which the test was able to categorize healthy pilots and people with known diagnoses or impairments (not always pilots).

Table 10 summarizes six studies that evaluated different assessments of memory. Most of these examined the relationship between memory and situational awareness. One study compared memory scores to imaging results that documented pathophysiological changes in brain tissue (McGuire, 2014). The study that is most directly related to the question this review hopes to answer reported the results of work by O'Donnell that was part of developing a battery called Neuropsychological Test Battery (NTB). The proposed battery included a memory retrieval task. While this study was well designed, it is difficult to draw a definitive conclusion about the memory task as the results differed for the memory test when it was included in different versions of the battery.

Table 10: Memory Assessments

Outcome	Assessment	Primary conclusions about test performance	Author, Year
Brain injury indicators (imaging)	Memory on MicroCog	U-2 pilots vs. controls: **adjusted p=0.036** U-2 pilots by level of WMH burden • WMH count**: adjusted p=0.030** • WMH volume: NS	McGuire, 2014[18]
Group comparison or classification	Sternberg memory retrieval	Version 2.0 (last version of the parent battery): Memory component **significantly discriminates between "pathologic" and healthy subjects** Inconsistent performance between version 1.1 and 2.0 of the test	O'Donnell, 1992[16]
Situational Awareness	Immediate/ Delayed Memory	• Short-term memory: not consistently correlated with SA • Total errors and SA: low correlation	Endsley, 1994[10]
Situational Awareness	Long-term working memory	**Predictor of SA in expert pilots**	Doane, 2003[8]
Situational Awareness	Building memory	All tested correlations: NS	Sulistyawati, 2011[20]
Situational Awareness	ATC Situation Recognition task	Total number of relevant cues reported**: Significant when spatial memory test included (adjusted R^2 = 0.46, p<0.005)** and improved prediction over spatial memory test only	Stokes, 1992[32]
Situational Awareness	Spatial memory test	Total number of relevant cues reported: **Significant predictor (adjusted R^2 = 0.32, p<0.01)**	Stokes, 1992[32]

ATC = air traffic control; NS = not statistically significant; SA = situational awareness; WMH = white matter hyperintensity. **Bold** = statistically significant

Table 11 contains the results reported on assessments of motor response. This was one of the least common domains represented in the included studies and the only significant findings were correlations in two studies; one with simulator performance and one with situational awareness.

Table 11: Motor Response

Outcome	Assessment	Primary test performance results	Author, Year
Group comparison or classification	Interval production test	Not significant in early versions and not included in subsequent versions	O'Donnell, 1992[16]
Group comparison or classification	Unstable tracking test	Significantly discriminates in early versions (1.0 and 1.1), but not in later version 2.0	O'Donnell, 1992[16]
Performance	Psychomotor task: single- and multitask	Both single- and multitask: r = NR, p=NS	Griffin, 1987[11]
Simulator Performance	Target hitting test	Discriminant analysis (pilots who made appropriate vs. inappropriate landing decision): NS Flight path deviations Correlation r = -0.39, NS (unadjusted) Regression analysis including landing decision: NS	Causse, 2010,[5] 2011a[6]
Simulator Performance	Target hitting test	**r = -0.45, R^2 = 0.20, p=0.050 for flight experience-partialed correlation**	Causse, 2011b[7]
Situational Awareness	Laser aiming task 2	Correlation value not reported, but **described as significant (p<0.05)**	Carretta, 1996[4]

NR = not reported; NS = not statistically significant
Bold = statistically significant

Six studies included assessments of perception. These are presented in Table 12. Most reported correlations with situational awareness scores or levels. One study provided some information on the relationship between perception and flight performance but this was in a study designed to compare pilots of manned and remotely operated aircraft (Barron, 2016). The one study that used simulator performance scores reported that perception was not significant in models that contained age and expertise (Taylor, 2007).

Table 12: Perception

Outcome	Assessment	Primary test performance results	Author, Year
Group comparison or classification	Manikin	Mean score in patients vs. pilots and healthy non pilots Speed: **significantly lower, p=0.01** Accuracy: NS	Kay, 1995[13] (Kay 1991/Clinical Studies I "FAA Phase B Study")[12]

Outcome	Assessment	Primary test performance results	Author, Year
Performance	Block counting	Manned aircraft pilots**: r = 0.054, p<0.05** Remotely piloted aircraft, r = -0.24, NS Corrected (adjusted) correlations reported but significance test not reported	Barron, 2016[31]
Performance	Block design test	Significantly correlated with all performance measures Significant independent predictor of all performance measures in regression models Correlations Aircraft route probe accuracy: r = 0.34, **p<0.001** Route recall accuracy: r = 0.38, **p<0.001** ATC message readback: r = 0.40, **p<0.001** Hierarchical regression models Model: sentence span + processing speed + block design + age + expertise Aircraft route probe accuracy: standardized beta = 0.24, **p<0.01** Route recall accuracy: standardized beta = 0.23, **p<0.01** ATC message readback: standardized beta = 0.28, **p<0.001**	Morrow, 2003[30]
Simulator Performance	Manikin	Level of expertise In model including age, expertise and age x expertise interaction. NS	Taylor, 2007[24]
Situational Awareness	Aerial Orientation Test	r = 0.150, p=NR	Endsley, 1994[10]
Situational Awareness	Cube comparison test	r = 0.353, p=NR	Endsley, 1994[10]
Situational Awareness	Cube comparison test	Level 1 SA: r = -0.09, p=NS Level 2 SA: r = 0.19, p=NS Level 3 SA: r = 0.27, p=NS controlled for flight hours	Sulistyawati, 2011[20]
Situational Awareness	Dot estimation	reaction time: r = -0.382, p=NR # correct: r = -0.415, p=NR	Endsley, 1994[10]
Situational Awareness	Encoding speed	reaction time - categorical subtest: r = -0.547, p=NR - physical subtest: r = -0.074, p=NR - name subtest: r = -0.295, p=NR total errors: r = -0.264, p=NR	Endsley, 1994[10]
Situational Awareness	Form board test	Level 1 SA: r = -0.45, p=NS Level 2 SA: r = -0.07, p=NS Level 3 SA: r = -0.30, p=NS controlled for flight hours	Sulistyawati, 2011[20]
Situational Awareness	Revised Minnesota Form Board Test	r = 0.317, p=NR	Endsley, 1994[10]
Situational Awareness	Perceptual speed	reaction time - subtest 5: r = -0.448, p=NR (shortest presentation time) - other subtests and total: r = -0.167 to 0.066, p=NR total errors: r = 0.366, p=NR	Endsley, 1994[10]
Situational Awareness	Group Embedded Figures test	r = 0.385, p=NR	Endsley, 1994[10]

Outcome	Assessment	Primary test performance results	Author, Year
Situational Awareness	Hidden figure test	Level 1 SA: **r = 0.58, p<0.05** Level 2 SA: r = -0.28, p=NS Level 3 SA: r = -0.01, p=NS controlled for flight hours	Sulistyawati, 2011[20]
Situational Awareness	Hidden patterns recognition	**adjusted R^2 = 0.46, p<0.05** Model: Spatial Memory incorrect response latency + Spatial Memory % correct previously unseen figures + Hidden Pattern Recognition % correct Model with spatial memory variables only: NS	Stokes, 1992[32]
Situational Awareness	Rotated hidden patterns	**adjusted R^2 = 0.34, p<0.05** Model: Risk-taking score + ATC situation recognition + Rotated hidden patterns Model with only situation recognition alone or with risk-taking: NS	Stokes, 1992[32]
Situational Awareness	Mental rotation ability	**adjusted R^2 = 0.54, p<0.05** Model: Spatial Memory test + ATC situation recognition + Mental rotation ability Model with Spatial Memory test + ATC situation recognition: NS	Stokes, 1992[32]

ATC = air traffic control; NR = not reported; NS = not statistically significant; SA = situational awareness
Bold = statistically significant

While situational awareness was used as an outcome is some studies, four studies evaluated situational awareness in terms of its correlations with simulator performance. These are summarized in Table 13.

Table 13: Situational Awareness

Outcome	Assessment	Primary test performance results	Author, Year
Simulator Performance	Situation Awareness	Correlation: situational awareness with simulator performance SA rated by observer F-15 pilots: r = 0.55, p=NR F-16 pilots: r = 0.60, p=NR SA rated by self-report F-15 pilots: r = 0.68, p=NR F-16 pilots: r = 0.85, p=NR	Venturino, 1990[42]
Simulator Performance	Situation Awareness Global Assessment Technique	Probability of kill Blue team (aggressors**): p=0.004** Red team (defenders): NS Pilot survival Blue Team (aggressors): NS Red Team (defenders): NS	Endsley, 1990[42]
Simulator Performance	Situation Awareness Global Assessment Technique	Pilot survival **Multiple R = 0.75, adjusted R^2 = 0.41, p<0.05** Model: Level 1 SA + Level 2 SA + Level 3 SA + overconfidence bias	Sulistyawati, 2011[20]

Outcome	Assessment	Primary test performance results	Author, Year
		Partial r for each Level: NS	
Simulator Performance	Situation Awareness Rating Scale	r = 0.56, p=NR	Bell, 1997[40]

NR = not reported; NS = not statistically significant; SA = situational awareness
Bold = statistically significant

Table 14 summarizes the results of evaluations of reasoning assessments. The results for reasoning are more consistent, with most studies reporting results that support a relationship between scores on a reasoning assessment and the selected outcome. The studies were nearly equally split between those that assess the test in terms of its ability to classify subjects as healthy or impaired or compare mean scores across groups and studies that used simulator performance.

Table 14: Reasoning

Outcome	Assessment	Primary test performance results	Author, Year
Brain injury indicators (imaging)	Reasoning / calculation domain	U-2 pilots vs. controls: **adjusted p=0.001** By level of WMH burden • WMH count: **adjusted p=0.044** • WMH volume: adjusted NS	McGuire, 2014[18]
Group comparison or classification	Arithmetic Test	Test discriminates between healthy subjects and those with a clinical diagnosis (**Version 2.0 significant**, earlier versions NS)	O'Donnell, 1992[16]
Group comparison or classification	Logical reasoning test nonsense syllogisms	Test discriminates between healthy subjects and those with a clinical diagnosis (**Version 2.0 significant**, earlier versions NS)	O'Donnell, 1992[16]
Group comparison or classification	Math test	Subjects with aviation performance issue, suspected neurologic and confirmed neurologic disorder scored worse than pilots referred for psychiatric or alcohol issues (no healthy group) **speed: p=0.015** **accuracy: p=0.000** **thruput: p=0.001** Similar result when adjusted for age.	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1)[13]
Group comparison or classification	Math test	Patients vs. pilots and healthy non-pilots Patients lower, ANCOVA **p=0.001**	Kay, 1995[13](Kay 1991/Clinical Studies I "FAA Phase B Study")[12]
Group comparison or classification	Pathfinder	Subjects with aviation performance issue, suspected neurologic and confirmed neurologic disorder scored worse than pilots referred for psychiatric or alcohol issues (no healthy group) **number speed: p=0.050** **letter speed: p=0.010** **letter thruput: p=0.030**	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1)

Outcome	Assessment	Primary test performance results	Author, Year
		combined speed: p=0.004 **combined thruput: p=0.010** Similar result when adjusted for age.	
Group comparison or classification	Pathfinder	Patients vs. pilots and healthy non-pilots Number speed: **p=0.01** Combined speed: **p=0.001** Difference score: **p<0.001** • Difference between number speed and combined speed: 1.3 seconds in patients vs. 0.6 seconds in pilots and 0.5 seconds in healthy non-pilots	Kay, 1995[13](Kay 1991/Clinical Studies I "FAA Phase B Study")[12]
Group comparison or classification	Verbal thinking test	Later versions (2.0 and 1.1) Test discrimination NS **First version (1.0) significant** (p=NR)	O'Donnell, 1992[16]
Performance	Verbal analogies	Manned aircraft and remote pilots All correlations NS	Barron, 2016[31]
Simulator Performance	Deductive reasoning test	**Significantly correlated with flight path deviation r = -0.63, p<0.01 and significant predictor in regression model (p=0.0083).** Discriminant analysis (pilots who made appropriate vs. inappropriate landing decision): NS	Causse, 2011a[6] Causse, 2010[5]
Simulator Performance	Math test	High workload condition: **r = -0.345, p=0.011** Low workload condition: NS	Tolton, 2014[27]
Simulator Performance	Pathfinder	High workload condition: **r = 0.416, p=0.002** Low workload condition: **r = 0.410, p=0.002**	Tolton, 2014[27]
Simulator Performance	Pathfinder	Exploratory ROC analysis Rate of decline in performance: p=NS	Yesavage, 2011[26]
Simulator Performance	Pathfinder	Initial performance: **significant for overall and 3 of 4 sub scores** Rate of decline in performance: NS In models including expertise, intra-individual variation, and executive function. Practice effects do not change the relationship between the assessment and simulator performance.	Kennedy, 2013[21] Kennedy, 2015[22]
Simulator Performance	Reasoning test	**r = -0.54, R^2 = 0.30, p=0.006** flight experience-partialed correlation	Causse, 2011b[7]
Simulator Performance	Wisconsin Card Sorting test	Discriminant analysis (pilots who made appropriate vs. inappropriate landing decision): p=NS Flight path deviations Correlation r = 0.25, p=NS Regression model: NS	Causse, 2011a Causse, 2010[5]
Simulator Performance	Wisconsin Card Sorting test	Test significantly correlated with correct decision **Corrected correlation r = 0.15, p=0.027**	Causse, 2011b[6]
Situational Awareness	Analytic test	Correlation r = 0.073, p=NR	Endsley, 1994[10]
Situational Awareness	Figure classification	Test correlation with all situational awareness levels: NS	Sulistyawati, 2011[20]
Situational Awareness	Following direction	Test correlation with all situational awareness levels: NS	Sulistyawati, 2011[20]

Outcome	Assessment	Primary test performance results	Author, Year
Situational Awareness	Math aptitude test	Correlation with Level 1 and 2 SA: NS Correlation with Level 3 Level 3 SA (projection of future status based on current situation): **r = 0.57, p<0.05**	Sulistyawati, 2011[20]
Situational Awareness	Logical reasoning test	Regression model: **adjusted R^2 = 0.60, p<0.005** Included level of certification and tests, when certification not considered: NS	Stokes, 1992[32]
Situational Awareness	Rapid serial classification 4-square	Correlation **r = NR, p<0.05** adjusted for flight experience	Carretta, 1996[4]
Situational Awareness	Raven's Advanced Progressive Matrices	Correlation: r = 0.243, p=NR	Endsley, 1994[10]

ANCOVA = analysis of covariance; NR = not reported; NS = not statistically significant; ROC = receiver operating characteristic; SA = situational awareness; WMH = white matter hyperintensity
Bold = statistically significant

Assessments of executive function evaluated in the included studies use tests that required judgements about how and when to shift attention or that required completing a complex task such as a maze. All but two of the studies report a significant relationship between the test and the outcome. The outcomes used included the entire ranges we found across topics: brain injury indications, classification, flight performance, simulator performance and situational awareness.

Table 15: Executive Function

Outcome	Assessment	Primary test performance results	Author, Year
Brain injury indicators	Shifting attention	Correlation with duration of post-traumatic amnesia: **r = 0.49, p<0.01**	Moore, 1995[19]
Group comparison or classification	Shifting attention	Difference across groups (pilots, healthy non pilots, and pilots/non pilot patients), after controlling for age **All measures significant. p<0.05**	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1)[13]
Performance	Shifting attention	Arrow direction thruput: Correlations significant for two pilot groups (TU-154 and IL-86) of one test **p<0.01** Discovery thruput: Correlations significant for TU-154 **p<0.01**; for IL-86 NS	Kay, 1995[13](Yakimovich 1994/Clinical Studies III "Actual flight errors")[14]
Simulator Performance	Shifting attention	Initial performance: **significant for overall and 1 of 4 sub scores (communications).** Other sub scores: NS Rate of decline in performance: NS In models including expertise, intra-individual variation, and executive function. Practice effects do not change the relationship between the assessment and simulator performance.	Kennedy, 2013[21] Kennedy, 2015[22]
Simulator Performance	Shifting attention	Correlation. **r = 0.43, p<0.05** In model with 3 levels of expertise: NS	Taylor, 2005[25]
Simulator Performance	Shifting attention	In model with 3 levels of expertise: NS	Taylor, 2007[24]
Simulator Performance	Shifting attention	<u>High workload</u> Instruction reaction speed: **r = -0.311, p=0.022** Discovery measures: NS <u>Low workload</u> Instruction reaction speed:: NS Discovery measures: r = 0.268, p=0.05	Tolton, 2014[27]
Simulator Performance	Shifting attention	Correlation **r = -0.324, p<0.05** Multiple regression **standardized beta = -0.216, p=0.032** Model included expertise, speed and working memory composite, and visual attention composite	Van Benthem, 2016[28]
Simulator Performance	Shifting attention	Rate of decline in performance: significantly lower **(p=0.008)** for pilots with higher baseline scores compared to pilots with lower scores of the same age. Initial baseline performance: NS	Yesavage, 2011[26]
Situational Awareness	Maze task	Correlation r = -0.354, p=NR	Endsley, 1994[10]
Situational Awareness	Maze tracking test	Regression model: **adjusted R^2 = 0.42, p<0.001** Included level of certification and tests, when certification not considered: NS	Stokes, 1992[32]

NR = not reported; NS = not statistically significant

Bold = statistically significant

Attention was the most frequently evaluated neuropsychological domain. We divided it into simpler tests presented in Table 16 and more complex attention tests in Table 17. The outcomes for tests of simple attention were simulator performance in several studies. The results varied for other outcomes, such as the ability to classify study subjects and situational awareness. The two studies that considered the relationship to brain injury outcomes were very different: one examined brain images and test performance and found no significant relationship (McGuire, 2014) while the other found that test scores were significantly correlated with the duration of post-traumatic amnesia (Moore, 1996). However, each of these approaches has limitations, imaging modality and resolution for the former, and accuracy of recall and reporting on the duration of amnesia for the latter.

Table 16: Attention - Simple

Outcome	Assessment	Primary test performance results	Author, Year
Brain injury indicators (imaging)	Timers 1 & 2	U-2 pilots vs. controls: NS U-2 pilots by level of WMH burden (count or volume): NS	McGuire, 2014[18]
Brain injury indicators (duration of post-traumatic amnesia)	CogScreen sub Tests	Visual Sequence Comparison **r = 0.53, p<0.01** Matching to Sample **r = 0.49, p<0.01** Divided Attention Test - premature responses: **r = 0.43, p<0.02** - speed: **r = 0.41, p<0.03**	Moore, 1995[19]
Group comparison or classification	Auditory Sequence Comparison\ Backward digit span Matching to Sample Visual Sequence Comparison	Patients vs. pilots and healthy non-pilots Auditory Sequence Comparison - speed: **p=0.001** - accuracy: NS Backward digit span: NS **Significantly worse performance in patients:** **Matching to Sample** **Visual Sequence Comparison**	Kay, 1995 [13](Kay 1991/Clinical Studies I "FAA Phase B Study")[12]
Group comparison or classification	Auditory Sequence Comparison Divided Attention Test Single Condition (as part of Dual Task Test) Matching to Sample Visual Sequence Comparison	**Significantly different mean scores across different groups**, all with potential clinical issues	Kay, 1995 [13](Clinical Studies II "Phase C Clinical Data" sample 1)
Group comparison or classification	Color-Word Test (modified Stroop) Dynamic memory test Spatial Processing Visual Monitoring	None included in final 2.0 Version in the course of developmental testing due to NS correlations in earlier versions.	O'Donnell, 1992[16]
Performance	Backward digit span Single Condition (as part of Dual Task Test)	Backward digit span TU-154 pilots: **r = -0.23, p<0.01** IL-86 pilots: **r = -0.46, p<0.01**	Kay, 1995[13] (Yakimovich 1994/Clinical Studies III

Outcome	Assessment	Primary test performance results	Author, Year
	Matching to Sample	Single Condition (as part of Dual Task Test) TU-154 pilots: **r = -0.26, p<0.01** IL-86 pilots: r = NR, p=NS Matching to Sample TU-154 pilots: **r = 0.24, p<0.01** IL-86 pilots: r = NR, p=NS	"Actual flight errors")[14]
Performance	Processing speed composite Sentence-span test	<u>Processing speed composite</u> - Significantly correlated with all performance measures - Significant independent predictor of all performance measures in regression models Correlations Aircraft route probe accuracy: r = 0.19**, p<0.05** Route recall accuracy: r = 0.25, **p<0.01** ATC message readback: r = 0.27, **p<0.001** Hierarchical regression models Model: sentence span + processing speed composite + block design + age + expertise • Aircraft route probe accuracy: standardized beta = -0.10, p=NS • Route recall accuracy: standardized beta = -0.05, p=NS • ATC message readback: standardized beta = -0.02, p=NS Model: processing speed composite + sentence span test • ATC readback task, accuracy only: adjusted R^2 = 0.001, p=NS <u>Sentence-span test</u> - Significantly correlated with all performance measures - Not a significant independent predictor of any performance measures in regression models Correlations Aircraft route probe accuracy: r = 0.40, **p<0.001** Route recall accuracy: r = 0.46, **p<0.001** ATC message readback: r = 0.42, **p<0.001** Hierarchical regression models Model: sentence span + processing speed composite + block design + age + expertise • Aircraft route probe accuracy: standardized beta = -0.34, **p<0.001** • Route recall accuracy: standardized beta = 0.38, **p<0.001** • ATC message readback: standardized beta = 0.31, **p<0.001** Model: processing speed composite + sentence span test • ATC readback task, accuracy only: adjusted R^2 = 0.097, **p<0.001**	Morrow, 2003[30]
Performance	Single Condition (as part of Dichotic listening)	All scores: NS	Griffin, 1987[11]

Outcome	Assessment	Primary test performance results	Author, Year
Simulator Performance	2-back Test Spatial Stroop	Discriminant analysis (pilots who made appropriate vs. inappropriate landing decision) 2-back test: **p<0.001** Spatial Stroop: NS Correlations and regression model for flight path deviations 2-back Test Regression model**: p=0.039** Correlations: r = -0.35, p=NS Partial correlation adjusted for flight experience **r = -0.41, R^2 = 0.17, p=0.022** Spatial Stroop Regression model: NS Correlation: NS Partial Correlation adjusted for flight experience: NS	Causse, 2010[5] and 2011a[6] Causse 2011b[6]
Simulator Performance	Single Condition (as part of Dual Task Test) Matching to Sample Visual Sequence Comparison	High Work Load Dual Task Test. NS Matching to Sample**. r = 0.355, p=0.009** Visual Sequence Comparison **r = -0.340, p=0.012** Low Work Load Dual Task Test. Mixed results Matching to Sample**. r = 0.523, p=0.001** Visual Sequence Comparison **r = -0.276, p=0.044**	Tolton, 2014[27]
Situational Awareness (SAGAT)	Auditory letter span test Number comparison test Visual number span test (modified)	Correlations for 3 levels of SA: all NS Adjusted for flight hours	Sulistyawati, 2011[20]
Situational Awareness	Continuous opposites XYZ assignment	**Correlations significant p<0.05** adjusted for flight experience (r value NR)	Carretta, 1996[4]
Situational Awareness	Internal timing Perceptual vigilance	Correlations r = -0.074, p=NR r = 0.041, p=NR	Endsley, 1994[10]

NR = not reported; NS = not statistically significant; ROC = receiver operating characteristic; SA = situational awareness; SAGAT = Situation Awareness Global Assessment Technique; WMH = white matter hyperintensity
Bold = statistically significant

Tests of complex attention included tests where the subject was expected to complete two tasks simultaneously, such as a simulated flight and math problems, or where attention is needed to recognize patterns. The idea is to test the abilities that may be needed in routine or emergency aircraft operation. The results for these types of tests are included in Table 17. In some cases the results suggest that a complex attention task can more accurately categorize someone who is impaired than a simpler test (e.g., Zhang, 1997). While the results are varied, most of the identified studies suggest there is a relationship between complex attention abilities and performance. The studies that did not report this relationship were designed to assess other factors such as age or expertise, but were included as they provided data on the assessment (Taylor, 2005; Taylor, 2007; and Yesavage, 2001).

Table 17: Attention - Complex

Outcome	Assessment	Primary test performance results	Author, Year
Brain injury indicators (imaging)	Attention/mental control	U-2 pilots vs. controls: NS U-2 pilots by level of WMH burden (count or volume): NS	McGuire, 2014[18]
Group comparison or classification	1. Difference score: Visual Sequence Comparison and Divided Attention Test 2. Symbol Digit Coding 3. Divided Attention Test	**1. Patients: 5% vs Pilots and healthy non pilots 2.4% p<0.05** **2. Number correct, immediate recall, delayed recall: all p<0.05.** Accuracy: NS **3. Significantly poorer performance in patients** **Sequence Comparison Speed: p=0.001** **Indicator Dual Speed: p=0.001**	Kay, 1995[13](Kay 1991/Clinical Studies I "FAA Phase B Study")[12]
Group comparison or classification	Divided Attention Test Dual Task Test (dual condition)	**Significantly different mean scores across different groups**, all with potential clinical issues	Kay, 1995[13] (Clinical Studies II "Phase C Clinical Data" sample 1)
Group comparison or classification	Dual task: flight simulator and math (dual and single conditions)	<u>Normal vs. hospitalized pilots</u> Satisfactory performance on single flight: NS Satisfactory performance on dual task: **p<0.05** Information processing speed of 2nd task w/o flight: NS Information processing speed of 2nd task with flight: NS Stress Index: NS Pilot Psychophysiological Reserve Capacity: **p<0.05** <u>Before vs. After Sleep Deprivation</u> Pilot Psychophysiological Reserve Capacity: **p<0.05**	Zhang, 1997[29]
Group comparison or classification	Symbol digit substitution test Trail Making Test	Discrimination: **Significant in all versions of battery** Except % correct in symbol digit substitution: NS	O'Donnell, 1992[16]
Performance	Dichotic listening task: multitask	**All correlations significant** dichotic listening task 1, multitask - offensive maneuvering: **r = 0.62, p<0.01** - kill-difference composite: **r = 0.49, p<0.05** dichotic listening task 2, multitask - overall air combat maneuvering score: **r = 0.49, p<0.05** - offensive maneuvering: **r = 0.60, p<0.01**	Griffin, 1987[11]
Performance	Divided Attention Test Dual Task Test (single and dual condition evaluated)	Divided Attention Test TU-154 pilots: **r = 0.38, p<0.01** IL-86 pilots: **r = 0.32, p<0.01** Dual Task Test accuracy - TU-154 pilots: **r = -0.31, p<0.01** - IL-86 pilots: r = NR, p=NS speed - TU-154 pilots: **r = 0.31, p<0.01** - IL-86 pilots: **r = 0.43, p<0.01** thruput - TU-154 pilots: **r = -0.34, p<0.01** - IL-86 pilots: **r = -0.51, p<0.01**	Kay, 1995[13](Yakimovich 1994/Clinical Studies III "Actual flight errors")[14]

Outcome	Assessment	Primary test performance results	Author, Year
Simulator Performance	Dual Task Test (single and dual conditions)	Correlation **r = -0.476, p<0.01** Multiple regression analysis **standardized beta = -0.307, p=0.011** Adjusted for expertise, speed and working memory composite, and cognitive flexibility composite	Van Benthem, 2016[28]
Simulator Performance	Dual Task Test (dual condition)	<u>High workload</u> Previous Number Dual Accuracy: **r = 0.364, p=0.007** Tracking dual boundary hits: NS Tracking dual error: NS <u>Low workload</u> Previous number dual accuracy**: r = 0.381, p=0.005** Tracking dual boundary hits: **r = -0.481, p=0.001** Tracking dual error: **r = -0.408, p=0.020**	Tolton, 2014[27]
Simulator Performance	Dual Task Test (dual condition)	Scores across 3 levels of expertise: NS	Taylor, 2007[24]
Simulator Performance	Dual Task Test (single and dual conditions) Symbol Digit Coding	Rate of decline in performance: NS	Yesavage, 2011[26]
Simulator Performance	Working memory span composite	Scores across 3 levels of expertise: NS	Taylor, 2005[25]
Situational Awareness	Attention sharing	Tracking task difficulty level: r = 0.717, p=NR 2-digit cancelation reaction time: r = -0.138, p=NR 8-digit cancelation reaction time: r = -0.250, p=NR	Endsley, 1994[10]
Situational Awareness	Scheduling 2 Time sharing 2	r = NR, **p<0.05** adjusted for flight experience	Carretta, 1996[4]
Situational Awareness	Working memory capacity	**Predictor of SA in novice pilots**	Doane, 2003[8]

NR = not reported; NS = not statistically significant; SA = situational awareness; WMH = white matter hyperintensity
Bold = statistically significant

Discussion

Limitations of the Evidence Base and Future Research Needs

There are numerous publications that address the topic of neuropsychological function in pilots that can reasonably be considered related to fitness to fly. However, a much smaller subset reports the results of research that answer the operational questions of interest and contribute to an evidence base to inform policy and practice.

The major limitation is the heterogeneity of the evidence. The studies we identified evaluated a wide range of tests and only a few tests were the subject of multiple studies. Furthermore, the tests that are currently required in the United States for the neuropsychological evaluation of pilots[45]were underrepresented in the literature. Lack of consistent, repeated results from several studies of the same test or assessment reduces our confidence in the findings. Because there are few studies of any test, it is more likely the next study could change our conclusions.

The studies were also heterogeneous in terms of how they evaluated the neuropsychological tests. Tests were evaluated using different outcomes such as the ability of the test to correctly classify a person as healthy or as having an impairment/disease; or to predict adequate or impaired performance during either real or simulated flight. There was also variability in the statistical methods used to make these determinations ranging from simple correlations to complex modeling. This makes it impossible to quantitatively combine results across studies using meta-analysis. Instead, we had to present the evidence and look for trends or patterns. This by definition requires subjective determinations and interpretation.

Additionally, some of the strongest and largest studies conducted to date were used to establish norms for tests. While this is a critical step in test development, further testing in additional samples is often needed to confirm the generalizability of test norms and repeated testing may be necessary to capture differences or changes in populations over time.

While there is agreement in the literature that certain neuropsychological domains are important to safely pilot an aircraft (e.g., working memory, executive function, and situational awareness) there appears to be less consensus in the field about what basic measures and approaches should be used in future studies, nor has an unequivocal research agenda been formulated that incorporates input from all stake-holders and identifies the key future research needed to move the field forward.

Other fields that have faced considerable controversy have been able to find common ground regarding approaches, testing, and data elements that should be included in future studies Consensus regarding basic features of useful studies in this field would greatly facilitate comparison of data across studies in the future.

Limitations of our Approach

Our search was limited to research on pilots. Evaluation of the literature in other safety sensitive professions such as maritime piloting, railroad, and other transportation workers was not included. This reflects our decision to focus time and resources on locating as many direct evaluations of pilots as possible. There were also some concerns among our expert advisory panel that data from other populations might have limited applicability to aviation. Additionally, while there is a robust literature on evaluation of pilot candidates, this literature focuses on a younger population, and emphasizes tests of aptitude, intelligence, and personality. These are less important to addressing our stated question of evaluating the fitness to fly of active pilots with known neurological conditions or who have questions arise regarding neuropsychological function.

While we included internet searches and searches of technical reporting to supplement our searches of citation databases of published articles, it is possible we did not locate all relevant studies and reports. Several of the frequently cited studies in this field were only available as conference presentations, technical reports, or theses. There may be other unpublished work that is relevant that we did not include. We mitigated this to the extent possible by asking technical experts and stakeholders who reviewed the draft to inform us about any potentially relevant studies not identified and either included or excluded.

Conclusion

After extensive searches, we were able to identify and compile the results from 28 studies reported in 39 publications or presentations. While this number suggests a moderate-sized body of evidence, the heterogeneity of tests evaluated and the research methods used limit our ability to draw conclusions with a high level of confidence. This is reflected in our rating of the strength of evidence as low. There is much variability in terms of tests assessed, and approaches to analyses in what are mostly smaller studies. Furthermore, the results are inconsistent and the evidence does not allow us to conclude which tests work best to detect or add to the assessment of different diagnoses or disorders.

Some tests underwent comparatively extensive evaluation as part of their development; however, we found that little research has been done to refine measures or follow-up after developmental studies to determine if tests continue to remain relevant to contemporary pilots, flying conditions, and policies. Given the importance of this topic, future research with larger numbers of subjects and rigorous approaches is warranted.

While the time and resources available limited the scope of this review to the question of what is known about neuropsychological tests used to evaluate individuals who are already pilots, data from studies of pilot candidates and trainees may be useful and could be included in future work. However, fundamental to our understanding about how to optimize the testing of pilots across their increasingly long lifespan of flying activities are longitudinal studies that assess the impact of aging, and commonly acquired disorders that impact neuropsychological function and may affect flying performance. Armed with this information it is more likely that decisions about fitness to fly will be made that appropriately balance safety concerns against the benefits of allowing pilots to fly for as long as possible.

In order to move this field forward and achieve the standards of evidence-based practice increasingly endorseed in other fields of medicine, it will be critical to develop a research agenda that prioritizes questions for further investigation and studies designs that can answer these questions. Protocols for doing this have been established[46] and can serve as a road-map for aerospace neuropsychology. Key elements that should be addressed in this research agenda include the following:

(1) Identify appropriate populations and minimum numbers needed for normative data collection. Included should be statements outlining the frequency with which normative data should be updated that is consistent with current best practices in the broader field of neuropsychology. Appropriate population norms which could be used in the evaluation of pilots should be identified whenever possible.
(2) Identify key tests, including tests used in the general population and aviation specific tests, that are most frequently used in the clinical assessment of pilots. Future studies should focus on the most frequently used clinical tests or new tests proposed to address gaps currently practice. This will strengthen the body of evidence supporting current approaches to aerospace neuropsychological testing.
(3) Consensus should be developed regarding the most appropriate study designs (including comparators, outcomes, and statistical approaches) needed to assure that future studies

are rigorous and directly address important quesitons. This will increase the likelihood that data can be compared and combined across studies This is especially important given the expense involved in conducting research studies and the relatively select population being studied.

(4) Identify other safety sensitive fields in which neuropsychological testing may be important and include representatives from these fields in the development of the research agenda. Given the long-history of work in this area within aviation, it is likely that aviation specific neuropsychology work may have relevance to other fields. Identifying opportunities to contribute to neuropsychological evaluation of individuals working in other safety sensitive areas will increase the impact of future work and increase potential funding opportunities.

(5) Identify existing unpublished data that could provide an important starting place for future research. Data is currently collected by airlines, airline pilot unions, and regulatory agencies. A set of guidelines based on what has been done in other fields would need to be developed that outline how such data could be protected, deidentified, and used in a fashion to strengthen the knowledge-base for aviation neuropsychological testing without compromising safeguards for pilot privacy and airline business models.

(6) Increase opportunities for the mentorship of junior neuropsychologists and researchers to encourage the development of research programs that sequentially build on prior studies and support the practice of the neuropsychological evaluation in safety sensitive professions such as aviation. Pathways for doing this are well-established in other fields[47]

References

1. Lezak M, Howieson D, Bigler E, Tranel D. *Neuropsychological Assessment.* 5th Edition ed. New York, NY: Oxford University Press; 2012.
2. Agency for Healthcare Research and Quality. *Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality.*
3. Eden J, Levit L, Berg A, et.al. *Finding what works in healthcare: Standards for systematic reviews.* Washington (DC): National Academic Press; 2011.
4. Carretta TR, Perry DC, Jr., Ree MJ. Prediction of situational awareness in F-15 pilots. *Int J Aviat Psychol.* 1996;6(1):21-41.
5. Causse M, Dehais F, Pastor J. Flight experience and executive functions predict flight simulator performance in general aviation pilots. 4th International Conference on Research in Air Transportation (ICRAT); 2010; Budapest, Hungary.
6. Causse M, Dehais F, Pastor J. Executive functions and pilot characteristics predict flight simulator performance in general aviation pilots. *Int J Aviat Psych.* 2011;21(3):217-234.
7. Causse M, Dehais F, Arexis M, Pastor J. Cognitive aging and flight performances in general aviation pilots. *Aging Neuropsychol Cogn.* 2011;18(5):544-561.
8. Doane S. *New measures of complex cognitive abilities: Relating memory processes to aviation flight situation awareness abilities.* Mississippi: Mississippi State University;2003.
9. Endsley MR. Predictive utility of an objective measure of situation awareness. *Proceedings of the Human Factors Society Annual Meeting.* 1990;34(1):41-45.

10. Endsley MR, Bolstad CA. Individual differences in pilot situation awareness. *Int J Aviat Psych.* 1994;4(3):241-264.
11. Griffin GR, Morrison TR, Amerson TL, Hamilton PV. Predicting air combat maneuvering (ACM) performance: Fleet fighter ACM readiness program grades as performance criteria. In. Command NMRaD, trans: Naval Aerospace Medical Research Laboratory; 1987.
12. Kay G, Horst R, Pakull B, Hordinsky J. Automated cognitive function assessment of civilian pilots. 62nd Annual Scientific Meeting of the Aerospace Medical Association; 1991; Cincinnati, Ohio, USA.
13. Kay GG. *CogScreen Aeromedical Edition professional manual.* 1995.
14. Yakimovich NV, Strogin GL, Govorushenko D, Schroeder D, Kay GG. Flight performance and CogScreen test battery in Russian pilots. Aerospace Medical Association; 1994.
15. Le Roux CGJ. *The development of a performance battery for mental and neurological screening in the aviation medical examiner's office*: Aerospace Medicine, Department of Community Medicine, Wright State University; 1988.
16. O'Donnell RD, Hordinsky JR, Madakasira S, Moise S, Warner D. A candidate automated test battery for neuropsychological screening of airmen: Design and preliminary validation. In: Administration FA, ed. Oklahoma City, Oklahoma1992.
17. McGuire SA, Boone GR, Sherman PM, et al. White matter integrity in high-altitude pilots exposed to hypobaria. *Aerosp Med Hum Perform.* 2016;87(12):983-988.
18. McGuire SA, Tate DF, Wood J, et al. Lower neurocognitive function in U-2 pilots: Relationship to white matter hyper intensities. *Neurology.* 2014;83(7):638-645.
19. Moore JL, Kay GG. Cog-Screen Aeromedical Edition in the assessment of the head injured military aviator. Aerospace Medical Association; 1995; Koln, Germany.
20. Sulistyawati K, Wickens CD, Chui YP. Prediction in situation awareness: Confidence bias and underlying cognitive abilities. *Int J Aviat Psych.* 2011;21(2):153-174.
21. Kennedy Q, Taylor J, Heraldez D, Noda A, Lazzeroni LC, Yesavage J. Intraindividual variability in basic reaction time predicts middle-aged and older pilots' flight simulator performance. *J Gerontol B Psychol Sci Soc Sci.* 2013;68(4):487-494.
22. Kennedy Q, Taylor J, Noda A, Yesavage J, Lazzeroni LC. The STEP model: Characterizing simultaneous time effects on practice for flight simulator performance among middle-aged and older pilots. *Psychol Aging.* 2015;30(3):699-711.
23. Taylor JL. "Relationship of CogScreen-AE to flight simulator performance and pilot age": Reply. *Aviat Space Environ Med.* 2000;71(11,Sect1):1166.
24. Taylor JL, Kennedy Q, Noda A, Yesavage JA. Pilot age and expertise predict flight simulator performance: A 3-year longitudinal study. *Neurology.* 2007;68(9):648-654.
25. Taylor JL, O'Hara R, Mumenthaler MS, Rosen AC, Yesavage JA. Cognitive ability, expertise, and age differences in following air-traffic control instructions. *Psychol Aging.* 2005;20(1):117-133.
26. Yesavage JA, Jo B, Adamson MM, et al. Initial cognitive performance predicts longitudinal aviator performance. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences.* 2011;66(4):444-453.
27. Tolton RG. *Relationship of individual pilot factors to simulated flight performance*: Occupational and Aviation Medicine, Thesis, University of Otago; 2014.

28. Van Benthem K, Herdman CM. Cognitive factors mediate the relation between age and flight path maintenance in general aviation. *Aviat Psych Appl Hum Fact.* 2016;6(2):81-90.
29. Zhang LM, Yu LS, Wang KN, Jing BS, Fang C. The psychophysiological assessment method for pilot's professional reliability. *Aviat Space Environ Med.* 1997;68(5):368-372.
30. Morrow D. Expertise, Cognitive Ability, and Age Effects on Pilot Communication. *Int J Aviat Psychol.* 2003;13(4):345-371.
31. Barron LG, Carretta TR, Rose MR. Aptitude and trait predictors of manned and unmanned aircraft pilot job performance. *Mil Psychol.* 2016;28(2):65-77.
32. Stokes A, Kemper K, March R. *Time-stressed flight decision making: A study of expert and novice aviators.* Idaho Falls, ID: Idaho National Engineering Laboratory;1992.
33. Basner M, Savitt A, Moore TM, et al. Development and validation of the cognition test battery for spaceflight. *Aerosp Med Hum Perform.* 2015;86(11):942-952.
34. Hyland D, Kay E. *Age 60 study, Part IV: Experimental evaluation of pilot performance.* Cherry Hill, New Jersey: Civil Aeromedical Institute Federal Aviation Administration;1994.
35. Kay G, Strongin G, Hordinsky J, Pakul B. Development of aviator norms for Cogscreen. Aerospace Medical Association Annual Scientific Meeting; 1993; Toronto, Canada.
36. Stokes AF, Banich MT, Elledge VC. Testing the tests--an empirical evaluation of screening tests for the detection of cognitive impairment in aviators. *Aviat Space Environ Med.* 1991;62(8):783-788.
37. DeVoll J. FAA Experience with Neuropsychological Testing for Airmen with Depression on SSRI Medications. *Aviat Space Environ Med.* 2013;Abstract #448(March):409.
38. Shull RN, Griffin GR. *Predicting F-14 air combat maneuvering (ACM) performance using automated battery of cognitive/psychomotor tests.* Pensacola, Florida1990. NAMRL-1356.
39. Kennedy Q, Taylor JL, Reade G, Yesavage JA. Age and expertise effects in aviation decision making and flight control in a flight simulator. *Aviat Space Environ Med.* 2010;81:489-497.
40. Bell HH, Waag WL. Using observer ratings to assess situational awareness in tactical air environments. In. Armstrong Laboratory USAF, trans: Armstrong Laboratory; 1997.
41. Waag WL, Houck MR. Tools for assessing situational awareness in an operational fighter environment. *Aviat Space Environ Med.* 1994;65(5 Suppl):A13-19.
42. Venturino M, Hamilton WL, Dvorchak S. Performance-based measures of merit for tactical situation awareness. Aerospace Medical Panel Symposium: Situational Awareness in Aerospace Operations; October 2-6, 1989, 1990; Copenhagen, Denmark.
43. Taylor JL, O'Hara R, Mumenthaler MS, Yesavage JA. Relationship of CogScreen-AE to flight simulator performance and pilot age. *Aviat Space Environ Med.* 2000;71(4):373-380.
44. Stokes AF, Belger A, Banich MT, Taylor H. Effects of acute aspartame and acute alcohol ingestion upon the cognitive performance of pilots. *Aviat Space Environ Med.* 1991;62(7):648-653.
45. Federal Aviation Administration. Guide for Aviation Medical Examiners. Decision Considerations Disease Protocols – Neurocognitive Impairment. https://www.faa.gov/about/office_org/headquarters_offices/avs/offices/aam/ame/guide/dec_cons/disease_prot/neurocog/ Accessed 10/28/2019.

46. Agency for Healthcare Research and Quality, 2010, Future Research Needs – Methods Research Series, https://effectivehealthcare.ahrq.gov/products/future-research-needs-methods/overview
Accessed 10/30/2019.
47. Katz S. Investing in the Early Stages of the Scientific Pipeline, 2015 https://www.niams.nih.gov/about/about-the-director/investing-early-stages-scientific, Accessed 10/27/2019.

Appendices

Appendix A. Research Team and Advisory Panel

Methods Team

Annette M. Totten, PhD	Methodologist
Eilis Boudreau, MD, PhD	Neurologist
Tamara P. Cheney, MD	Research Associate
Cynthia Davis-O'Reilly, BS	Project Manager

Pacific Northwest Evidence-based Practice Center
Oregon Health & Science University, Portland, OR

Advisory Panel Members

In designing the study questions and methodology at the outset and during the preparation of this report, the Pacific Northwest Evidence-based Practice Center (EPC) consulted an Advisory Panel. This panel included members representing technical and content experts and potential end users of the research. Broad expertise and perspectives were sought. Divergent and conflicting opinions are common and perceived as healthy scientific discourse that results in a thoughtful, relevant systematic review. Advisory Panel members were not involved in the analysis of the evidence or the writing of the report. Therefore, study questions, design, methodological approaches, and/or conclusions do not necessarily represent the views of individual Advisory Panel members.

Advisory Panel members were requested to disclose any financial conflicts of interest and any other relevant business or professional conflicts of interests. None were disclosed. The list of Advisory Panel members follows:

Fred Bonato, PhD, Aerospace Medical Association
Nick Caplan, PhD, Aerospace Medicine Systematic Review Group
James Devoll, MD, Office of Aerospace Medicine, Federal Aviation Administration
Jay Dorothy, Allied Pilots Association
John Hastings, MD, Neurologist
Pete Lewis, Allied Pilots Association
Ed Miles, PhD, Clinical Psychologist Allied Pilots Association
Muriel Lezak, PhD, Neuropsychologist Emeritus Professor Oregon Health and Science University
Scott Rossow, D.O. CFII, Aerospace Medical Certification Division, Federal Aviation Administration
Andrew Winnard, PhD, Aerospace Medicine Systematic Review Group
Mona Nasser, DDS Cochrane Methods Group

Appendix B. Inclusion and Exclusion Criteria

Full Text Paper Inclusion/Exclusion Codes: Reasons for full text paper inclusion or exclusion

Inclusion
1 = Include in the report **99** = complex/unclear

Exclusion
2 = Background or discussion paper only, no data for evidence, but pull for review
3 = Incorrect population (non-human, general population, not pilots or similar profession)
3a = Pilot selection or training
3b = Similar professions (e.g. air traffic controllers, other transportation workers)
4 = No included assessment or no assessment (not an assessment of neuropsychological function)
5 = No included comparison or no comparison
6 = Does not have an included outcome (e.g., feasibility only, descriptive, no outcome)
7 = Wrong timing or setting
8 = Condition not included (e.g. transient states such as sleep deprivation/hypoxia)
9 = Excluded study design (nonsystematic reviews, evaluation of hypothetical uses or needs assessment
10 = Wrong publication type (opinion, editorial, letter, guideline document not used for background)
11 = Wrong years (studies published before 1980)
12 = Not in English, but may be relevant

	Include	**Exclude**
Populations	Pilots: general aviation, commercial (including air carrier/transport), military Any age As needed (maybe used as Indirect) • Similar professions (e.g. air traffic controllers, transportation workers)	General population Other professions Trainee Evaluation Pilot Selection
Tests	Any Neuropsychological assessment strategy, test or test battery Including but not limited to the list below	Tests that evaluate characteristics that are not neurocognition or neuropsychological; Tests of intelligence, aptitude, personality Examples: • Intelligence • Personality Type or Traits • Fatigue • Sleep Deprivation • Aptitude • Risk Aversion
Comparators	Comparisons to: gold standard, other tests, generally accepted values or cut offs or pilot performance measures including simulator performance	Reports that describe a test/assessment but provide no evaluation
Conditions	Include but not limited to the following that could affect neuropsychological function: • Stroke • Alcoholism • Substance use disorder	Do not include transient states related to the environment or are not the result of medical conditions. Examples: • Sleep Deprivation

	Include	Exclude
	• Head Trauma or TBI • Encephalitis • Multiple Sclerosis • HIV • Myocardial infarction • Chronic medication effects • Other suspected or developmental conditions that impact neurological function	• Hypoxia • Anxiety • Fatigue • Human Factors • Intoxication/acute substance or medication effects
Outcomes	Tests that are evaluated • Specificity • Sensitivity • Predictive Utility • Odds Ratio • AUROC • Reliability	Description of tests, but no evaluation
Timing	Evaluation of active pilots If needed: • Pre-employment screening • Assessment during training	
Setting	Outpatient	All others
Study Design and Publication Type	Evaluations or comparative studies, including cohorts and trials. Cohorts can be prospective or retrospective. Pre/post studies and cross section studies can be included.	Nonsystematic reviews, commentaries, or letters. Evaluations of hypothetical situations/synthetic data Descriptions of tests/assessments with no evaluation Case reports
Years	1980 to present	1979 and earlier
Language	English or other languages as long as an English abstract is available	Non English with no English abstract

Appendix C. Search Strategy

Bibliographic databases
Databases Searched – Medline, Psycinfo, Scopus
Database: Ovid MEDLINE(R) <1946 to January Week 4 2018>
Search Strategy:
---

1 Aerospace Medicine/ or exp Aviation/
2 Accidents, Aviation/pc [Prevention & Control]
3 astronauts/ or pilots/
4 (airline or aircraft or aircrew or aviator or (air* adj2 pilot*)).ti,ab.
5 exp Cognition Disorders/di, pc [Diagnosis, Prevention & Control]
6 exp Stress, Psychological/di, pc [Diagnosis, Prevention & Control]
7 exp Neuropsychological Tests/
8 exp Psychomotor Performance/
9 (cognition or cognitive or neurocognitive).ti,ab.
10 cogscreen.ti,ab,kw.
11 exp "reproducibility of results"/ or exp "sensitivity and specificity"/
12 di.fs.
13 or/1-4
14 13 and (5 or 6)
15 13 and (7 or 8 or 9)
16 15 and (11 or 12)
17 10 or 14 or 16
18 Aviation Space & Environmental Medicine.jn.
19 18 and (5 or 6)
20 (7 or 8 or 9) and (11 or 12)
21 18 and 20
22 17 or 19 or 21

FAA Aerospace Medicine Technical Reports
https://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/
Dates: through 9/13/18; two searches
Search number: term(s)

1. Cognitive Assessment
2. Cogscreen

OpenGrey
http://www.opengrey.eu
Search date: through 2/20/19; five Searches
Search number: term(s)

1. 'discipline(01* OR 15* OR 22*) AND cognit* lang:"en"'
2. 'discipline(01* OR 15* OR 22*) AND neuropsych* lang:"en"'
3. 'discipline:(06N AND 01*)'
4. 'aviat* cognit* lang:"en"'
5. 'discipline:(06J* OR 06N*) discipline:(01* OR 22*)'

National Technical Reports Library
https://ntrl.ntis.gov/NTRL/
Search date: through 3/19/19; one advanced search
Search strategy:

- Advanced Search: '(airline OR aircraft OR aviation OR aircrew OR pilots) AND (cognitive OR cognition OR neurocognitive)'
- Filtered on option: "Only documents with full text"
- Additional Fields:
    - Keywords (separate search terms): 'evaluation' OR 'assess' OR 'test' or 'dysfunction'
    - Date Published: 1980 to 2019
    - Refine by: Keyword: 'Pilots'

Royal Air Force (RAF) UK: Literature Database
Not available at this time

Appendix D. Included Studies

1. Barron, L.G., T.R. Carretta, and M.R. Rose, *Aptitude and trait predictors of manned and unmanned aircraft pilot job performance.* Military Psychology, 2016. **28**(2): p. 65-77 DOI: 10.1037/mil0000109.
2. Basner, M., et al., *Development and validation of the cognition test battery for spaceflight.* Aerospace Medicine and Human Performance, 2015. **86**(11): p. 942-952.
3. Bell, H.H. and W.L. Waag, *Using observer ratings to assess situational awareness in tactical air environments*. 1997, Armstrong Laboratory.
4. Carretta, T.R., D.C. Perry, Jr., and M.J. Ree, *Prediction of situational awareness in F-15 pilots.* International Journal of Aviation Psychology, 1996. **6**(1): p. 21-41.
5. Causse, M., et al., *Cognitive aging and flight performances in general aviation pilots.* Aging, Neuropsychology, and Cognition, 2011. **18**(5): p. 544-561.
6. Causse, M., F. Dehais, and J. Pastor, *Flight experience and executive functions predict flight simulator performance in general aviation pilots*, in *4th International Conference on Research in Air Transportation (ICRAT)*. 2010: Budapest, Hungary.
7. Causse, M., F. Dehais, and J. Pastor, *Executive functions and pilot characteristics predict flight simulator performance in general aviation pilots.* The International Journal of Aviation Psychology, 2011. **21**(3): p. 217-234 DOI: http://dx.doi.org/10.1080/10508414.2011.582441.
8. DeVoll, J., *FAA Experience with Neuropsychological Testing for Airmen with Depression on SSRI Medications.* Aviation Space & Environmental Medicine, 2013. **Abstract #448**(March): p. 409.
9. Doane, S., *New measures of complex cognitive abilities: Relating memory processes to aviation flight situation awareness abilities*. 2003, Mississippi State University: Mississippi.
10. Endsley, M.R., *Predictive utility of an objective measure of situation awareness.* Proceedings of the Human Factors Society Annual Meeting, 1990. **34**(1): p. 41-45 DOI: 10.1177/154193129003400110.
11. Endsley, M.R. and C.A. Bolstad, *Individual differences in pilot situation awareness.* The International Journal of Aviation Psychology, 1994. **4**(3): p. 241-264 DOI: http://dx.doi.org/10.1207/s15327108ijap0403_3.
12. Griffin, G.R., et al., *Predicting air combat maneuvering (ACM) performance: Fleet fighter ACM readiness program grades as performance criteria*. 1987, Naval Aerospace Medical Research Laboratory.
13. Hyland, D. and E. Kay, *Age 60 study, Part IV: Experimental evaluation of pilot performance*, J.D. Deimler and I. Hilton Systems, Editors. 1994, Civil Aeromedical Institute Federal Aviation Administration: Cherry Hill, New Jersey. p. 1-33.
14. Kay, G., et al., *Automated cognitive function assessment of civilian pilots*, in *62nd Annual Scientific Meeting of the Aerospace Medical Association*. 1991, Aviation, Space, and Environmental Medicine: Cincinnati, Ohio, USA.
15. Kay, G., et al., *Development of aviator norms for Cogscreen*, in *Aerospace Medical Association Annual Scientific Meeting*. 1993: Toronto, Canada.
16. Kay, G.G., *CogScreen Aeromedical Edition professional manual*. 1995.
17. Kennedy, Q., et al., *Intraindividual variability in basic reaction time predicts middle-aged and older pilots' flight simulator performance.* Journals of Gerontology Series B-

Psychological Sciences & Social Sciences, 2013. **68**(4): p. 487-94 DOI: https://dx.doi.org/10.1093/geronb/gbs090.
18. Kennedy, Q., et al., *The STEP model: Characterizing simultaneous time effects on practice for flight simulator performance among middle-aged and older pilots.* Psychology and Aging, 2015. **30**(3): p. 699-711 DOI: http://dx.doi.org/10.1037/pag0000043.
19. Kennedy, Q., et al., *Age and expertise effects in aviation decision making and flight control in a flight simulator.* Aviation Space & Environmental Medicine, 2010. **81**: p. 489-97.
20. Le Roux, C.G.J., *The development of a performance battery for mental and neurological screening in the aviation medical examiner's office*, in *Aerospace Medicine, Department of Community Medicine*. 1988, Wright State University. p. 101.
21. McGuire, S.A., et al., *White matter integrity in high-altitude pilots exposed to hypobaria.* Aerospace Medicine and Human Performance, 2016. **87**(12): p. 983-988 DOI: http://dx.doi.org/10.3357/AMHP.4585.2016.
22. McGuire, S.A., et al., *Lower neurocognitive function in U-2 pilots: Relationship to white matter hyper intensities.* Neurology, 2014. **83**(7): p. 638-645 DOI: 10.1212/WNL.0000000000000694.
23. Moore, J.L. and G.G. Kay, *Cog-Screen Aeromedical Edition in the assessment of the head injured military aviator*, in *Aerospace Medical Association*. 1995: Koln, Germany.
24. Morrow, D., *Expertise, Cognitive Ability, and Age Effects on Pilot Communication.* International Journal of Aviation Psychology, 2003. **13**(4): p. 345-371.
25. O'Donnell, R.D., et al., *A candidate automated test battery for neuropsychological screening of airmen: Design and preliminary validation*, F.A. Administration, Editor. 1992: Oklahoma City, Oklahoma.
26. Shull, R.N. and G.R. Griffin, *Predicting F-14 air combat maneuvering (ACM) performance using automated battery of cognitive/psychomotor tests*. 1990: Pensacola, Florida.
27. Stokes, A., K. Kemper, and R. March, *Time-stressed flight decision making: A study of expert and novice aviators*. 1992, Idaho National Engineering Laboratory: Idaho Falls, ID. p. 1-42.
28. Stokes, A.F., M.T. Banich, and V.C. Elledge, *Testing the tests--an empirical evaluation of screening tests for the detection of cognitive impairment in aviators.* Aviation Space & Environmental Medicine, 1991. **62**(8): p. 783-8.
29. Sulistyawati, K., C.D. Wickens, and Y.P. Chui, *Prediction in situation awareness: Confidence bias and underlying cognitive abilities.* The International Journal of Aviation Psychology, 2011. **21**(2): p. 153-174 DOI: http://dx.doi.org/10.1080/10508414.2011.556492.
30. Taylor, J.L., *"Relationship of CogScreen-AE to flight simulator performance and pilot age": Reply.* Aviation, Space, and Environmental Medicine, 2000. **71**(11,Sect1): p. 1166.
31. Taylor, J.L., et al., *Pilot age and expertise predict flight simulator performance: A 3-year longitudinal study.* Neurology, 2007. **68**(9): p. 648-654 DOI: http://dx.doi.org/10.1212/01.wnl.0000255943.10045.c0.
32. Taylor, J.L., et al., *Cognitive ability, expertise, and age differences in following air-traffic control instructions.* Psychology & Aging, 2005. **20**(1): p. 117-33.

33. Tolton, R.G., *Relationship of individual pilot factors to simulated flight performance*, in *Occupational and Aviation Medicine*. 2014, Thesis, University of Otago. p. 170.
34. Van Benthem, K. and C.M. Herdman, *Cognitive factors mediate the relation between age and flight path maintenance in general aviation.* Aviation Psychology and Applied Human Factors, 2016. **6**(2): p. 81-90 DOI: http://dx.doi.org/10.1027/2192-0923/a000102.
35. Venturino, M., W.L. Hamilton, and S. Dvorchak, *Performance-based measures of merit for tactical situation awareness*, in *Aerospace Medical Panel Symposium: Situational Awareness in Aerospace Operations*. 1990, NATO Advisory Group for Aerospace Research & Development: Copenhagen, Denmark.
36. Waag, W.L. and M.R. Houck, *Tools for assessing situational awareness in an operational fighter environment.* Aviation Space & Environmental Medicine, 1994. **65**(5 Suppl): p. A13-9.
37. Yakimovich, N.V., et al., *Flight performance and CogScreen test battery in Russian pilots*, in *Aerospace Medical Association*. 1994.
38. Yesavage, J.A., et al., *Initial cognitive performance predicts longitudinal aviator performance.* The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences, 2011. **66**(4): p. 444-453 DOI: http://dx.doi.org/10.1093/geronb/gbr031.
39. Zhang, L.M., et al., *The psychophysiological assessment method for pilot's professional reliability.* Aviation Space & Environmental Medicine, 1997. **68**(5): p. 368-72.

Appendix E. Excluded Studies

The following list contains studies that we reviewed at the full text level and excluded. A reason for exclusion is at the end of each citation.

1. Angelici A, Baker S, Cimrmancic MA, et al. The age 60 rule. Aviat Space Environ Med. 2004;75(8):708-15. Exclusion: Non-systematic Review
2. Banich MT, Stokes A, Elledge VC. Neuropsychological screening of aviators: A review. Aviat Space Environ Med. 1989 Apr;60(4):361-6. PMID: 1989-27821-001. Exclusion: Non-Systematic Review
3. Bautsch HS, Narayanan S, McNeese MD. Development and evaluation of a cognitive model of human-performance in fighter aircraft. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics; 1997. pp. 2109-13. Exclusion: No included assessment
4. Benline TA, French J, Poole E. Anti-emetic drug effects on pilot performance: Granisetron vs. ondansetron. Aviat Space Environ Med. 1997 Nov;68(11):998-1005. PMID: 1997-43825-003. Exclusion: No assessment
5. Bennett G. Medical-cause accidents in commercial aviation. Eur Heart J. 1992 Dec;13 Suppl H:13-5. PMID: 1493817. Exclusion: Descriptive
6. Bennett W, Jr., Schreiber BT, Andrews DH. Developing competency-based methods for near-real-time air combat problem solving assessment. Computers in Human Behavior. 2002 Nov;18(6):773-82. doi: http://dx.doi.org/10.1016/S0747-5632%2802%2900030-4. PMID: 2002-06400-014. Exclusion: No included assessment
7. Berry M. Aeromedical Advisory - The Limits of SimulationNovember/December. FAA Safety Briefing. 2017 November/December 2017:5. Exclusion: Editorial
8. Boer LC, Harsveld M, Hermans PH. The selective-listening task as a test for pilots and air traffic controllers. Mil Psychol. 1997;9(2):137-49. PMID: 11540404. Exclusion: Pilot selection or training
9. Borghini G, Aricò P, Di Flumeri G, et al. Avionic technology testing by using a cognitive neurometric index: A study with professional helicopter pilots. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS; 2015. In: 2015-November. pp. 6182-5. Exclusion: No included assessment
10. Bower EA, Moore JL, Moss M, et al. The effects of single-dose fexofenadine, diphenhydramine, and placebo on cognitive performance in flight personnel. Aviat Space Environ Med. 2003 Feb;74(2):145-52. PMID: 12602446. Exclusion: Not included condition
11. Burke E, Hobson C, Linsky C. Large sample validations of three general predictors of pilot training success. Int J Aviat Psychol. 1997;7(3):225-34. PMID: 11540941. Exclusion: Pilot selection or training
12. . Flight simulator fidelity considerations for total air line pilot training and evaluation. AIAA Modeling and Simulation Technologies Conference and Exhibit; 2001. In. Exclusion: No included assessment
13. Byrne MD, Kirlik A. Using computational cognitive modeling to diagnose possible sources of aviation error. Int J Aviat Psych. 2005;15(2):135-55. Exclusion: No included assessment

14. Caldwell JA, Stephens RL, Carter DJ, et al. Effects of 2 mg and 4 mg atropine sulfate on the performance of U.S. Army helicopter pilots. Aviat Space Environ Med. 1992 Oct;63(10):857-64. PMID: 1993-04756-001. Exclusion: Condition not included
15. Callister JD, King RE, Retzlaff PD. Cognitive assessment of USAF pilot training candidates. Aviat Space Environ Med. 1996 Dec;67(12):1124-9. PMID: 8968475. Exclusion: Pilot selection or training
16. Cannavo R, Conti D, Di Nuovo A. Computer-aided assessment of aviation pilots attention: Design of an integrated test and its empirical validation. Applied Computing and Informatics 2016 Google Search August 2018;12:16-26. Exclusion: No included comparison
17. Carretta TR. Basic Attributes Tests (BAT) System: Development of an Automated Test Battery for Pilot Selection Air Force Human Resources Laboratory. 1987. PMID: AD-A185 649. Exclusion: Pilot Selection or training
18. Carretta TR. Development and Validation of the Test of Aviation Skills (TBAS) Air Force Research Laboratory. 2005. PMID: ADA442563. Exclusion: Pilot selection or training
19. Causse M, Chua Z, Matton N. Can efficiency and prefrontal activity during a working memory task predict pilots' flight performance. Open Archive Toulouse Archive Ouverte. 2018. Exclusion: Pilot selection or training
20. Causse M, Matton N, Del Campo N. Cognition and piloting performance: offline and online measurements. HFES Europe Chapter Conference; 2012 Toulouse. Exclusion: No included assessment
21. Chappelle WL, Ree MJ, Barto EL, et al. Joint Use of the MAB-II and MicroCog for Improvement in the Clinical and Neuropsychological Screening & Aeromedical Waiver Process of Rated USAF Pilots. AFRL-SA-BR-TR-2010-0002. 2010. Exclusion: Pilot selection or training
22. Connemann BJ. Donepezil and flight simulator performance: effects on retention of complex skills. Neurology. 2003 Sep 09;61(5):721; author reply PMID: 12963782. Exclusion: Letter to the Editor
23. Corona BM, Fiedler ER. Potential paradigm for assessments of biomedical technologies in the operational environment. Aviat Space Environ Med. 2007 May;78(5 Suppl):B245-51. PMID: 17547325. Exclusion: No included assessment
24. Correia R. The use of psychotropic medications and fitness to fly. Pilot mental health assessment and support: A practitioner's guide. New York, NY: Routledge/Taylor & Francis Group; US; 2017:229-44. Exclusion: Descriptive
25. Cuevas J, Osterich H. Cross-cultural evaluation of the booklet version of the Category Test. Int J Clin Neuropsych. 1990;12(3-4):187-90. PMID: 1992-29737-001. Exclusion: Descriptive
26. Damos DL. Some Considerations in the Design of a Computerized Human Information Processing Battery Naval Aerospace Medical Research Laboratory. 1987. PMID: AD-A199 491. Exclusion: Descriptive
27. Damos DL. KSAOs for military pilot selection: A review of the literature. DAS-2011-01. 2011. PMID: AFCAPS-FR-2011-0003. Exclusion: Non-systematic review
28. Damos DL, Gibb GD. Development of a computer-based Naval aviation selection test battery Naval Aerospace Medical Research Laboratory. 1986. PMID: AD-A179 997. Exclusion: No included outcome

29. Debring CE, Van Gorp WG, Stuck AE. Early detection of cognitive decline in higher cognitvely functioning older adults: Sensitivity and specificity of a neuropsychological screening battery. Neuropsychology. 1994;8(1):31-7. Exclusion: Incorrect population
30. D'Oliveira TC. Dynamic spatial ability & ability to coordinate information: Predictive contributions of dynamic spatial ability and the ability to coordinate information for air traffic controller and pilot selection. Int J Applied Aviat Stud. 2003;3(2):227-41. PMID: 2005-00246-003. Exclusion: Pilot selection or training
31. Durso FT, Bleckley MK, Dattel AR. Does situation awareness add to the validity of cognitive tests? Hum Factors. 2006;48(4):721-33. PMID: 17240720. Exclusion: Incorrect population
32. Eissfeldt H, Grasshoff D, Hasse C, et al. Aviator 2030 - Ability requirements in future ATM Systems II: Simulations and experiments. DLR-FB-2009-28. 2009. PMID: ISSN: 1434-8454. Exclusion: No included assessment
33. Elwood RW. MicroCog: assessment of cognitive functioning. Neuropsychol Rev. 2001 Jun;11(2):89-100. PMID: 11572473. Exclusion: Background
34. Emery B. Neurocognitive predictors of flight performance of successful solo flight students. Dissertation Abstracts International: Section B: The Sciences and Engineering. 2012;73(3-B):1868. PMID: 2012-99180-030. Exclusion: Pilot selection or training
35. Endsley MR. Toward a theory of situation awareness in dynamic systems. Human Factors: The Journal of the Human Factors and Ergonomics Society. 1995;37(1):32-64. doi: 10.1518/001872095779049543. Exclusion: Descriptive/Background
36. Engelberg AL, Gibbons HL, Doege TC. A review of the medical standards for civilian airmen: Synopsis of a two-year study. JAMA. 1986;255(12):1589-99. Exclusion: Descriptive
37. Eonta SE, Carr W, McArdle JJ, et al. Automated neuropsychological assessment metrics: repeated assessment with two military samples. Aviat Space Environ Med. 2011 Jan;82(1):34-9. PMID: 21235103. Exclusion: Incorrect population
38. Fatolitis P, Jentsch F, Hancock P, et al. Initial validation of novel performance-based measures: Mental rotation and psychomotor ability Naval Aerospace Medical Research Laboratory. 2010. PMID: ADA529481. Exclusion: Incorrect population
39. Feng C, Wanyan X, Liu S, et al. Dynamic prediction model of situation awareness in flight simulation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); 2018. 10906 LNAI. pp. 115-26. Exclusion: Incorrect population
40. Fiedler ER, Orme DR, Mills W. Assessment of head-injured aircrew: Comparison of FAA and USAF procedures Civil Aerospace Medical Institute Federal Aviation Administration. DOT/FAA/AM-01/11. Oklahoma City, Oklahoma: 2001. Exclusion: Descriptive
41. Flynn CF, King RE. Using computerized neuropsychological testing to assess aviator skills Armstrong Laboratory. AL/AO-TR-1994-0174. 1994. PMID: ADA293227. Exclusion: Pilot selection or training
42. Flynn CF, Sipes WE, Grosenbach MJ, et al. Top performer survey: Computerized psychological assessment in aircrew. Aviat Space Environ Med. 1994 May;65(5, Sect 2, Suppl):A39-A44. PMID: 1994-35723-001. Exclusion: No included assessment

43. Fosha R, Blalock LD, Kass S. Cognitive constructs and psychomotor tracking. Advances in Intelligent Systems and Computing. 2017;484:663-75. doi: 10.1007/978-3-319-41682-3_56. Exclusion: No included assessment
44. Fracker ML. Measures of situation awareness: Review and future directions. AL-TR-1991-0128. 1991. PMID: AD-A262 672. Exclusion: Incorrect population
45. Froom P, Benbassat J, Gross M, et al. Air accidents, pilot experience, and disease-related inflight sudden incapacitation. Aviat Space Environ Med. 1988 Mar;59(3):278-81. PMID: 3355485. Exclusion: No included assessment
46. Gaetan S, Dousset E, Marqueste T, et al. Cognitive workload and psychophysiological parameters during multitask activity in helicopter pilots. Aerosp Med Hum Perform. 2015 Dec;86(12):1052-7. doi: http://dx.doi.org/10.3357/AMHP.4228.2015. PMID: 2015-56007-001. Exclusion: Background
47. Georgemiller R, Machizawa S, Young KM, et al. Neuropsychological assessment of decision making in alcohol-dependent commercial pilots. Aviat Space Environ Med. 2013 Sep;84(9):980-5. PMID: 2013-42166-005. Exclusion: No included outcome
48. Gerathewohl SJ. Psychophysiological effects of aging: Developing a functional age index for pilots: I. A survey of the pertinent literature. FAA-AM-77-6. 1977. Exclusion: Descriptive
49. Giannakoulas G, Katramados A, Melas N, et al. Acute effects of nicotine withdrawal syndrome in pilots during flight. Aviat Space Environ Med. 2003 Mar;74(3):247-51. PMID: 2003-02329-002. Exclusion: Condition not included
50. Gibellato MG, Moore JL, Selby K, et al. Effects of lovastatin and pravastatin on cognitive function in military aircrew. Aviat Space Environ Med. 2001 Sep;72(9):805-12. PMID: 2001-18663-001. Exclusion: No Included Comparison
51. Giffin WC, Rockwell TH. Computer-aided testing of pilot response to critical in-flight events. Hum Factors. 1984;26(5):573-81. Exclusion: No included outcome
52. Gilliland K, Schlegel RE. A laboratory model of readiness-to-perform testing. I: Learning rates and reliability analyses for candidate testing measures. DOT/FAA/AM-97/5. 1997. Exclusion: Incorrect population
53. Goeters KM, Maschke P, Eissfeldt H. Ability requirements in core aviation professions: Job analyses of airline pilots and airtraffic controllers. In: Goeters KM, ed Aviation psychology: Practice and research. London: Routledge; 2004. Exclusion: No included assessment
54. Gontar P, Hoermann H-J. Interrater reliability at the top end: Measures of pilots' nontechnical performance. Int J Aviat Psych. 2015 Oct;25(3-4):171-90. doi: http://dx.doi.org/10.1080/10508414.2015.1162636. PMID: 2016-27609-004. Exclusion: No included assessment
55. Gordon S, Goren C, Carmon E, et al. Cognitive evaluation of Israeli Air Force pilot cadets. Aerosp Med Hum Perform. 2017 Apr;88(4):392-8. doi: http://dx.doi.org/10.3357/AMHP.4677.2017. PMID: 2017-26551-001. Exclusion: Pilot selection or training
56. Gray R, Gaska J, Winterbottom M. Relationship between sustained, orientated, divided, and selective attention and simulated aviation performance: Training & pressure effects. Journal of Applied Research in Memory and Cognition. 2016;5(1):34-42. doi: 10.1016/j.jarmac.2015.11.005. Exclusion: Background

57. Griffin GR. Predicting naval aviator flight training performance using multiple regression and an artificial neural network. Int J Aviat Psych. 1998;8(2):121-35. doi: http://dx.doi.org/10.1207/s15327108ijap0802_3. PMID: 1998-04780-003. Exclusion: Pilot selection or training
58. Griffin GR, Shull RN. Predicting F/A-18 fleet replacement squadron performance using an automated battery of performance-based tests Naval Aerospace Medical Research Laboratory. NAMRL-1354. 1990. PMID: AD-A223 899. Exclusion: No included outcome
59. Gualtieri CT, Johnson LG. Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. Archives of Clinical Neuropsychology. 2006;21(7):623-43. Exclusion: Incorrect population
60. Hamilton HC. Airline pilots in recovery from alcoholism: A quantitative study of cognitive change. Dissertation Abstracts International: Section B: The Sciences and Engineering. 2016;77(3-B(E)) PMID: 2016-37853-181. Exclusion: No included assessment
61. Hardy DJ, Parasuraman R. Cognition and flight performance in older pilots. Journal of Experimental Psychology: Applied. 1997;3(4):313-48. Exclusion: No included comparison
62. Hardy DJ, Satz P, D'Elia LF, et al. Age-related group and individual differences in aircraft pilot cognition. Int J Aviat Psych. 2007;17(1):77-90. doi: http://dx.doi.org/10.1207/s15327108ijap1701_5. PMID: 2007-01565-005. Exclusion: No included outcome
63. Harris J, Wiggins M, Morrison B, et al. Differentiating cognitive complexity and cognitive load in high and low demand flight simulation tasks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); 2015. In: 8433. pp. 133-50. Exclusion: No included outcome
64. Harris JM, Wiggins MW. Evaluating cognitive competence: Does eye movement behavior represent the missing piece of the puzzle? Proceedings of the Human Factors and Ergonomics Society; 2008. 3. pp. 2077-81. Exclusion: No incuded outcome
65. Harris JM, Wiggins MW, Taylor S, et al. Performance and cognition in dynamic environments: The development of a new tool to assist practitioners. Multimodal Safety Management and Human Factors: Crossing the Borders of Medical, Aviation, Road and Rail Industries. 2012:159-68. Exclusion: No included assessment
66. Heil MC. Air traffic control specialist age and cognitive test performance Office of Aviation Medicine, Federal Aviation Administration. DOT/FAA/AM-99/23. 1999. Exclusion: Incorrect population
67. Hess DW, Kennedy CH, Hardin RA, et al. Attention Deficit/Hyperactivity Disorder and Learning Disorders. In: Kennedy C, Moore J, eds. Mil Neuropsych. New York, United States: Springer Publishing Company; 2010:199-226. Exclusion: Descriptive
68. Hoffman C, Hoffman A. The role of assessment in pilot selection. 2016:40-60. Exclusion: Pilot selection or training
69. Hoffman CC, Hoffman KP, Kay GG. The role that cognitive ability plays in CRM. NATO RTO Human Factors and Medicine Panel (HFM) Symposium. Edinburgh, United Kingdom: RTO Meeting Proceedings 4; 1998. p. 37-1 - -22. Exclusion: No included assessment

70. Hoft S, Pecena Y. Behaviour-oriented evaluation of aviation personnel: An assessment center approach. Aviation Psychology:practice and research. 2004:153-70. Exclusion: Pilot selection or training
71. Houston RC. Pilot personnel selection. Applications of interactionist psychology: Essays in honor of Saul B Sells. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc; US; 1988:291-316. Exclusion: No included outcome
72. Hsu CK, Lin SC, Li WC. Visual movement and mental-workload for pilot performance assessment. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015;9174:356-64. doi: 10.1007/978-3-319-20373-7_34. Exclusion: No included assessment
73. Hunter DR. Aviator selection. Military personnel measurement: Testing, assignment, evaluation. New York, NY, England: Praeger Publishers; England; 1989:129-67. Exclusion: Pilot selection or training
74. Hyland D. Age 60, Part II: Airline pilot age and performance-A review of the scientific literature. DOT/FAA/AM-94/21. Cherry Hill, New Jersery: 1994. Exclusion: Descriptive
75. Imhoff DL, Levine JM. Perceptual-motor and cognitive performance task battery for pilot selection. AFHRL-TR-80-27. 1981. PMID: AD-A094 317 81-1-30-033. Exclusion: Pilot selection or training
76. Intano GP, Howse WR, Lofaro RJ. The selection of an experimental test battery for aviator cognitive, psychomotor abilities and personal traits U.S. Army Research Institute. 1991. PMID: AD-A231 887. Exclusion: No included comparison
77. Jagathesan T, Obrien MD. Transient global amnesia and its aeromedical implications. Aviat Space Environ Med. 2012;83(6):565-9. doi: 10.3357/ASEM.2714.2012. Exclusion: Condition not included
78. James M, Green R. Airline pilot incapacitation survey. Aviat Space Environ Med. 1991;62(November):1068-72. Exclusion: No included assessment
79. Jipp M. Reaction times to consecutive automation failures: A function of working memory and sustained attention. Hum Factors. 2016 Dec;58(8):1248-61. doi: http://dx.doi.org/10.1177/0018720816662374. PMID: 2016-55500-009. Exclusion: Incorrect population
80. Johansson K, Lundberg C. Assessment of fitness to drive, possession of professional drivers' license, possession of firearms, and pilot's certificate in clients with dementing conditions. Competence assessment in dementia. New York, NY: Springer Publishing Co; US; 2008:93-101. Exclusion: Incorrect population
81. John MS, Kobus DA, Morrison JG, et al. Overview of the DARPA augmented cognition technical integration experiment. International Journal of Human-Computer Interaction. 2004;17(2):131-49. doi: http://dx.doi.org/10.1207/s15327590ijhc1702_2. PMID: 2004-16864-002. Exclusion: Incorrect population
82. Jones DR. Aerospace Psychiatry. In: Davis JR, ed Fundamentals of Aerospace. 2008:406-24. Exclusion: Pilot selection or training
83. Kane RL, Kay GG. Computerized assessment in neuropsychology: a review of tests and test batteries. Neuropsychology Review. 1992;3(1):1-117. Exclusion: Descriptive
84. Katchen MS. Synopsis of the many phases of H.I.V. infection. Aviat Space Environ Med. 1990 Aug;61(8):763-4. PMID: 2400386. Exclusion: Letter to the Editor

85. Katz LC. Finding the "Right Stuff": Development of an Army Aviator Selection Instrument U.S. Army Research Institute for the Behaviorial and Social Sciences. 0704-0188. Fort Rucker, Alabama: 2006. Exclusion: No included outcome
86. Kautz MA, Thomas ML, Caldwell J. Considerations of pharmacology on fitness for duty in the operational environment. Aviat Space Environ Med. 2007 May;78(5, Sect II, Suppl):B107-B12. PMID: 2007-07068-015. Exclusion: No included outcome
87. Kay E, Hillman D, Hyland D, et al. Age 60 Study, Part III: Consolidated Database Experiments Final Report. 1994. Exclusion: No included assessment
88. Kay G. Effects of aging on aviation-related cognitive functioning in commercial airline pilots. [Abstract 55]. Aviat Space Environ Med. 2001;72:241. Exclusion: No included comparison
89. Kay G, Hordinsky JR, Pakull B. Neuropsychological assessment of aviators: A comparison of traditional and computer-based approaches. 63rd Annual Scientific Meeting of the Aerospace Medical Association; 1992 Miami Beach, Florida, USA. Aviation, Space, and Environmental Medicine. Exclusion: Wrong publication type
90. Kay GG. Guidelines for the psychological evaluation of air crew personnel. Occup Med. 2002 Apr-Jun;17(2):227-45, iv. PMID: 11872438. Exclusion: No included assessment
91. Kay GG. Aviation neuropsychology. Aeromedical psychology. Aldershot, England: Ashgate Publishing Ltd; England; 2013:239-68. Exclusion: Background
92. Kelly MJ. PERFORMANCE MEASUREMENT DURING SIMULATED AIR-tO-AIR COMBAT. Hum Factors. 1988;30(4):495-506. Exclusion: No included outcome
93. Kennedy RS, Fowlkes JE, Lilienthal MG. Postural and performance changes following exposures to flight simulators. Aviat Space Environ Med. 1993 Oct;64(10):912-20. PMID: 1994-07762-001. Exclusion: No included assessment
94. King RE. Assessing Aviators for Personality Pathology with the Millon Clinical Multiaxial Inventory (MCMI). Aviat Space Environ Med. 1994;65:227-31. Exclusion: No included assessment
95. King RE, Barto E, Ree MJ, et al. Compilation of Pilot Cognitive Ability Norms. AFRL-SA-WP-TR-2012-0001. Air Force Research Labratory: 2011. Exclusion: Background
96. King RE, Flynn CF. Defining and measuring the "right stuff": neuropsychiatrically enhanced flight screening (N-EFS). Aviat Space Environ Med. 1995 Oct;66(10):951-6. PMID: 8526831. Exclusion: No included outcome
97. Landau DA, Azaria B, Fineman R, et al. Long-term survivors of childhood malignancies - Aeromedical dilemmas and implications. Aviat Space Environ Med. 2006;77(12):1266-70. Exclusion: No assessment
98. Larcher J, Veronneau S, DeJohn C. In-Flight Medical Incapacitation Research. [Abstract 359]. Aviat Space Environ Med. 2001;72:306. Exclusion: No included assessment
99. Lehenbauer LP. An investigation of the construct-related and criterion-related validity of cogscreen-aeromedical edition. Dissertation Abstracts International: Section B: The Sciences and Engineering. 2003;64(4-B):1931. PMID: 2003-95020-166. Exclusion: Pilot Selection
100. Leirer VO, Yesavage JA, Morrow DG. Marijuana, aging, and task difficulty effects on pilot performance. Aviat Space Environ Med. 1989 Dec;60(12):1145-52. PMID: 2604668. Exclusion: Condition not included

101. Li G, Baker SP, Grabowski JG, et al. Age, Flight Experience, and Risk of Crash Involvement in a Cohort of Professional Pilots. American Journal of Epidemiology. 2003;157(10):874-80. doi: 10.1093/aje/kwg071. Exclusion: No included assessment
102. Lieberman HR, Kramer FM, Montain SJ, et al. Field assessment and enhancement of cognitive performance: development of an ambulatory vigilance monitor. Aviat Space Environ Med. 2007 May;78(5 Suppl):B268-75. PMID: 17547328. Exclusion: No included assessment
103. . Study on mental attributes of aged test pilots. Lecture Notes in Electrical Engineering; 2018. In: 456. Exclusion: No included outcome
104. Longridge T, Bürki-Cohen J, Go TH, et al. Simulator fidelity considerations for training and evaluation of today's airline pilots. 2001. Exclusion: Background
105. Manzey D. Monitoring of mental performance during spaceflight. Aviat Space Environ Med. 2000 Sep;71(9 Suppl):A69-75. PMID: 10993313. Exclusion: Incorrect population
106. Mapou RL, Kay GG, Rundell JR, et al. Measuring performance decrements in aviation personnel infected with the human immunodeficiency virus. Aviat Space Environ Med. 1993 Feb;64(2):158-64. PMID: 1993-22284-001. Exclusion: Non-systematic review
107. Mapou RL, Rundell JR, Kay GG, et al. Relating cognitive function to military aviator performance in early HIV infection. Vaccine. 1993;11(5):555-9. Exclusion: Descriptive
108. Maroco J, Bartolo-Ribeiro R. Selection of Air Force pilot candidates: A case study on the predictive accuracy of discriminant analysis, logistic regression, and four neural network types. Int J Aviat Psych. 2013 Apr;23(2):130-52. doi: http://dx.doi.org/10.1080/10508414.2013.772837. PMID: 2013-12824-003. Exclusion: Pilot selection or training
109. Martin-Saint-Laurent A, Lavernhe J, Casano G, et al. Clinical aspects of inflight incapacitations in commercial aviation. Aviat Space Environ Med. 1990 Mar;61(3):256-60. PMID: 2317181. Exclusion: No included assessment
110. Martinussen M. Psychological measures as predictors of pilot performance: a meta-analysis. Int J Aviat Psychol. 1996;6(1):1-20. PMID: 11539171. Exclusion: Pilot Selection
111. Martinussen M. Pilot selection: An overview of aptitude and ability assessment. Pilot mental health assessment and support: A practitioner's guide. New York, NY: Routledge/Taylor & Francis Group; US; 2017:23-39. Exclusion: Descriptive
112. Martinussen M, Torjussen T. Pilot selection in the Norwegian Air Force: A validation and meta-analysis of the test battery. Int J Aviat Psych. 1998;8(1):33-45. Exclusion: Pilot training or selection
113. Maschke P, Oubaid V, Pecena Y. How do astronaut candidate profiles differ from airline pilot profiles? Results from the 2008/2009 ESA astronaut selection. Aviat Psych Appl Hum Fact. 2011;1(1):38-44. doi: http://dx.doi.org/10.1027/2192-0923/a00006. PMID: 2011-11428-007. Exclusion: Pilot selection or training
114. Matton N, Vautier S, Raufaste E. Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. Intelligence. 2009 Jul-Aug;37(4):412-21. doi: http://dx.doi.org/10.1016/j.intell.2009.03.011. PMID: 2009-08754-011. Exclusion: Pilot selection or training
115. Matton N, Vautier S, Raufaste E. Test-specificity of the advantage of retaking cognitive ability tests. International Journal of Selection and Assessment. 2011 Mar;19(1):11-7.

doi: http://dx.doi.org/10.1111/j.1468-2389.2011.00530.x. PMID: 2011-03309-002. Exclusion: Pilot selection or training
116. McFadden KL. Predicting pilot-error incidents of US airline pilots using logistic regression. Appl Ergon. 1997 Jun;28(3):209-12. PMID: 9414359. Exclusion: No included assessment
117. McGuire SA, Marsh RW, Sowin TW, et al. Aeromedical decision making and seizure risk after traumatic brain injury: Longitudinal outcome. Aviat Space Environ Med. 2012;83(2):140-3. doi: 10.3357/ASEM.3104.2012. Exclusion: No included assessment
118. McKenna FP. Effects of unattended emotional stimuli on color-naming performance. Current Psychological Research & Reviews. 1986 Spr;5(1):3-9. doi: http://dx.doi.org/10.1007/BF02686591. PMID: 1987-06267-001. Exclusion: Pilot Selection
119. . Assessing behaviour of cognitive agents in a flight simulator with fighter pilots. 2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings; 2017. In. Exclusion: No included assessment
120. Milke RM, Becker JT, Lambrou P. The Effects of Age and Practice on Aviation-Relevant Concurrent Task Performance Federal Aviation Administration. 1999. Exclusion: Incorrect population
121. Miller EN, Selnes OA. Aviation safety and asymptomatic HIV-1 infection. Aviat Space Environ Med. 1993 Nov;64(11):1059-60. PMID: 8280042. Exclusion: Letter
122. Mironov AD, Nakvasin AY. Development and flight research of non-intrusive airborne system of pilot psychological status monitoring State Scientific Centre of the Russian Federation Gromov Flight Research Institute 2012. Exclusion: No included assessment
123. Mohler SR, Nicholson A, Harvey P, et al. The use of antihistamines in safety-critical jobs: A meeting report. Current Medical Research and Opinion. 2002;18(6):332-7. doi: 10.1185/030079902125000877. Exclusion: Non-Systematic Review
124. Momen N. The Effects of Alternative Input Devices and Repeated Exposures on the Test of Basic Aviation Skills (TBAS) Performance. Mil Med. 2009;174(12):1282-6. PMID: ADA512274. Exclusion: Pilot selection or training
125. Morrow D, Leirer V, Yesavage J. The influence of alcohol and aging on radio communication during flight. Aviat Space Environ Med. 1990 Jan;61(1):12-20. PMID: 2302121. Exclusion: Condition not included
126. Morrow D, Yesavage J, Leirer V, et al. The time-course of alcohol impairment of general aviation pilot performance in a Frasca 141 simulator. Aviat Space Environ Med. 1993 Aug;64(8):697-705. PMID: 8368982. Exclusion: Not included condition
127. Muller A, Petru R, Angerer P. Cognitive demands and the relationship between age and workload in apron control. Aviat Space Environ Med. 2011 Jan;82(1):26-33. doi: http://dx.doi.org/10.3357/ASEM.2797.2011. PMID: 2010-26861-003. Exclusion: Incorrect population
128. Mumenthaler MS, Benowitz NL, Taylor JL, et al. Nicotine deprivation and pilot performance during simulated flight. Aviat Space Environ Med. 2010;81(7):660-4. doi: 10.3357/ASEM.2701.2010. Exclusion: Condition not included
129. Mumenthaler MS, Yesavage JA, Taylor JL, et al. Psychoactive drugs and pilot performance: A comparison of nicotine, donepezil, and alcohol effects. Neuropsychopharmacology. 2003 Jul;28(7):1366-73. doi:

http://dx.doi.org/10.1038/sj.npp.1300202. PMID: 2003-06456-020. Exclusion: No included comparison
130. Nicholson A. The Neurosciencecs and the Practice of Aviation Medicine. Boca Raton, Florida: CRC Press Taylor and Francis Group; 2011. Exclusion: No included outcome
131. Nicholson AN, Turner C. Anti-Diuretic for Operational Aircrew: Effects of Desmopressin on Urine Flow, Cognition, and Sleepiness. Aviat Space Environ Med. 2005 Aug;76(8):760-5. PMID: 2005-08707-003. Exclusion: Condition not included
132. Nidos A, Kontostavlos SS, Roussos P, et al. Aerospace neuropsychology: Exploring the construct of psychological and cognitive interaction in the 100 most fatal civil aviation accidents through multidimensional scaling. 2018. (vol. 586). Exclusion:
133. Jensen RS, Neumeister D, eds. Expertise in Aeronautical Decision Making: A Cognitive Skill Analysis. Seventh International Symposium on Aviation Psychology; 1993 April 26-29, 1993; Columbus, Ohio, United States. The International Journal of Aviation Psychology; 1. Exclusion: No included assessment
134. Olson RB. An analysis of student progress in beginning flight training: Performance prediction, performance measurement, and performance improvement. Dissertation Abstracts International: Section B: The Sciences and Engineering. 2003 Mar;63(9-B):4405. PMID: 2003-95006-276. Exclusion: Pilot selection or training
135. Orme DR. The path back to the cockpit: FAA and USAF procedures for head-injured aviators. Human Factors and Aerospace Safety. 2003;3(4):353-64. PMID: 2004-15229-004. Exclusion: No included outcome
136. Ostoin SD. An Assessment of the Performance-Based Measurement Battery (PBMB), the Navy's Psychomotor Supplement to the Aviation Selection Test Battery (ASTB): Naval Postgraduate School; 2007. Exclusion: No included comparison
137. Pattyn N, Mairesse O, Cortoos A, et al. Cardiac reactivity and preserved performance under stress: Two sides of the same coin? Int J Psychophysiol. 2014 Jul;93(1):30-7. doi: http://dx.doi.org/10.1016/j.ijpsycho.2013.03.008. PMID: 2013-13224-001. Exclusion: Pilot selection or training
138. Paul MA, MacLellan M, Gray G. Motion-Sickness Medications for Aircrew: Impact on Psychomotor Performance. Aviat Space Environ Med. 2005 Jun;76(6,Sect1):560-5. PMID: 2005-06418-004. Exclusion: Condition not included
139. Prince C, Salas E. Situation assessment for routine flight and decision making. International Journal of Cognitive Ergonomics. 1998;1(4):315-24. Exclusion: No included outcome
140. Razsolov NA, Krapivnitskaya TA, Khashba BG. Intelligence quotients of pilots in civil aviation with atherosclerotic changes in the brain vessels. Human Physiology. 2012;38(7):751-2. doi: 10.1134/S0362119712070213. Exclusion: No included assessment
141. Rebok GW, Li G, Baker SP, et al. Self-rated changes in cognition and piloting skills: A comparison of younger and older airline pilots. Aviat Space Environ Med. 2002;73(5):466-71. Exclusion: No included assessment
142. Ree MJ, Carretta TR, Teachout MS. Role of ability and prior knowledge in complex training performance. J Appl Psychol. 1995 Dec;80(6):721-30. doi: http://dx.doi.org/10.1037/0021-9010.80.6.721. PMID: 1996-18597-001. Exclusion: Pilot selection or training

143. Reid M, Parkinson L, Gibson R, et al. Memory complaint questionnaire performed poorly as screening tool: validation against psychometric tests and affective measures. J Clin Epidemiol. 2012 Feb;65(2):199-205. doi: https://dx.doi.org/10.1016/j.jclinepi.2011.06.006. PMID: 21889305. Exclusion: Incorrect population
144. Retzlaff PD, Callister JD, King RE. The Computerized Neuropsychological Evaluation of US Air Force Pilots: Clinical Procedures and Data-Based Decision. AL/AO-TR-1996-0107. 1996. Exclusion: Background
145. Retzlaff PD, Callister JD, King RE. Clinical procedures for the neuropsychological evaluation of U.S. Air Force pilots. Mil Med. 1999 Jul;164(7):514-9. PMID: 10414068. Exclusion: Pilot selection or training
146. Retzlaff PD, King RE, Callister JD. USAF Pilot Training Completion and Retention: A Ten Year Follow-Up on Psychological Testing. AL/AO-TR-1995-0124. 1995. Exclusion: No included comparison
147. Ridgway J. Construct validity through facet analysis: Scheduling tests do not necessarily measure scheduling ability. Journal of Occupational Psychology. 1980 Dec;53(4):253-63. doi: http://dx.doi.org/10.1111/j.2044-8325.1980.tb00032.x. PMID: 1981-06787-001. Exclusion: Incorrect population
148. Ross SM. Cognitive function following exposure to contaminated air on commercial aircraft: A case series of 27 pilots seen for clinical purposes. Journal of Nutritional and Environmental Medicine. 2008;17(2):111-26. doi: 10.1080/13590840802240067. Exclusion: No comparison
149. Ross SM. Assessing cognitive function in airline pilots: The importance of neuropsychological assessment. Pilot mental health assessment and support: A practitioner's guide. New York, NY: Routledge/Taylor & Francis Group; US; 2017:116-30. Exclusion: No included outcome
150. Roth W-M. Flight examiners' methods of ascertaining pilot proficiency. Int J Aviat Psych. 2015 Oct;25(3-4):209-26. doi: http://dx.doi.org/10.1080/10508414.2015.1162642. PMID: 2016-27609-006. Exclusion: No included assessment
151. Russo MB, Stetz MC, Thomas ML. Monitoring and predicting cognitive state and performance via physiological correlates of neuronal signals. Aviat Space Environ Med. 2005 Jul;76(7 Suppl):C59-63. PMID: 16018331. Exclusion: Descriptive
152. Salive ME. Evaluation of aging pilots: Evidence, policy, and future directions. Mil Med. 1994 Feb;159(2):83-6. PMID: 1994-32072-001. Exclusion: No included outcome
153. Sataloff RT, Hawkshaw M, Kutinsky J, et al. The aging physician and surgeon. Ear Nose Throat J. 2016 Apr-May;95(4-5):E35-48. PMID: 27140028. Exclusion: No included outcome
154. Schroeder DJ, Harris HC, Jr., Collins WE, et al. Some Performance Effects of Age and Low Blood Alcohol Levels on a Computerized Neuropsychological Test. DOT/FAA/AM-95/7. 1995. Exclusion: Incorrect population
155. Selnes OA, Miller EN. Asymptomatic HIV-1 infection and aviation safety. Aviat Space Environ Med. 1993 Feb;64(2):172-5. PMID: 8431194. Exclusion: Letter to the Editor
156. Shappell S, Bartosh B. Use of a Commercially Available Flight Simulator During Aircrew Performance Testing Department of the Navy, Aerospace Med Research Laboratory. 1991. Exclusion: Condition not included

157. Shephard JM, Kosslyn SM. The minicog rapid assessment battery: developing a "blood pressure cuff for the mind". Aviat Space Environ Med. 2005 Jun;76(6 Suppl):B192-7. PMID: 15943212. Exclusion: No included outcome
158. Shull R. Performance of Marine AV-8B (Harrier) Pilots on a Cognitive/Psychomotor Test Battery: Comparison and Prediction: apps.dtic.mil; 1990. Exclusion: No included comparison
159. Smith JK, Caldwell JA. Methodology for Evaluating the Simulator Flight Performance of Pilots Air Force Research Laboratory. Brooks City-Base Texas: 2004. Exclusion: Condition not included
160. Steinkraus LW, Rayman RB, Butler WP, et al. Aeromedical decision making-It may be time for a paradigm change. Aviat Space Environ Med. 2012 Oct;83(10):1006-7. doi: http://dx.doi.org/10.3357/ASEM.3406.2012. PMID: 2012-27403-002. Exclusion: Commentary
161. Stetz MC, Thomas ML, Russo MB, et al. Stress, mental health, and cognition: a brief review of relationships and countermeasures. Aviat Space Environ Med. 2007 May;78(5 Suppl):B252-60. PMID: 17547326. Exclusion: Descriptive
162. Stokes A, Belger A, Banich M, et al. Effects of alcohol and chronic aspartame ingestion upon performance in aviation relevant cognitive tasks. Aviation, space, and …. 1994. Exclusion: Condition not included
163. Tarnowski A, Terelak J. Double Trait Assessment Test Battery for Air Force Pilots Department of Aviation Psychology, Air Force Institute of Aviation Medicine. 1999. PMID: ADA362216. Exclusion: Descriptive
164. Taylor J, Mumenthaler MS, Kraemer HC, et al. Longitudinal Study of Older Small-Aircraft Pilots: Changes in Cogscreen-AE Performance. 2001. Exclusion: No included outcome
165. Tenney YJ, Adams MJ, Pew RW, et al. A Principled Approach to the Measurement of Situation Awareness in Commercial Aviation. NASA-CR-4451. 1992. PMID: Contract NAS1-18788, BBN-7451. Exclusion: Background
166. Thomas JR, Schrot J. Naval Medical Research Institute Performance Assessment Battery (NMRI PAB) Documentation Naval Medical Research Institute. 1988. PMID: AD-A201 654. Exclusion: Descriptive
167. Thompson WT, Orme DR, Zazeckis TM. Neuropsychological Evaluation of Aviators: Need for Aviator-Specific Norms? SAM-FE-BR-TR-2004-0001. 2004. Exclusion: Pilot selection or training
168. Trenite DG, Vermeiren R. The impact of subclinical epileptiform discharges on complex tasks and cognition: Relevance for aircrew and air traffic controllers. Epilepsy & Behavior. 2005 Feb;6(1):31-4. doi: http://dx.doi.org/10.1016/j.yebeh.2004.10.005. PMID: 2005-01650-008. Exclusion: No included assessment
169. Truszczynska A, Lewkowicz R, Truszczynski O, et al. Back pain and its consequences among Polish air force pilots flying high performance aircraft. International Journal of Occupational Medicine and Environmental Health. 2014 Apr;27(2):243-51. doi: http://dx.doi.org/10.2478/s13382-014-0254-z. PMID: 2014-36814-011. Exclusion: No included assessment
170. Tsang PS. Assessing Cognitive Aging in Piloting. In: Tsang PS, Vidulich MA, eds. Principles and Practice of Aviation Psychology. 2003:425-64. Exclusion: Descriptive

171. Uhlarik J, Comerford DA. A Review of Situation Awareness Literature Relevant to Pilot Surveillance Functions. DOT/FAA/AM-02/3. 2002. PMID: 98F80691. Exclusion: Descriptive
172. Uitdewilligen S, de Voogt A. Cognitive skill correlates of the automated pilot selection system. Human Factors and Aerospace Safety. 2006;6(4):333-44. PMID: 2008-02258-006. Exclusion: Pilot selection or training
173. Vacchiano C, Moore J, Rice GM, et al. Fexofenadine effects on cognitive performance in aviators at ground level and simulated altitude. Aviat Space Environ Med. 2008 Aug;79(8):754-60. doi: http://dx.doi.org/10.3357/ASEM.2212.2008. PMID: 2008-11065-001. Exclusion: Not included condition
174. Valk PJL, Simons R, Jetten AM, et al. Cognitive performance effects of bilastine 20 mg during 6 hours at 8000 ft cabin altitude. Aerosp Med Hum Perform. 2016;87(7):622-7. doi: 10.3357/AMHP.4522.2016. Exclusion: Not included condition
175. Valk PJL, Van Roon DB, Simons RM, et al. Desloratadine Shows No Effect on Performance during 6 h at 8,000 ft Simulated Cabin Altitude. Aviat Space Environ Med. 2004;75(5):433-8. Exclusion: Not included condition
176. Veldhuis M, Matton N, Vautier S. Using item response theory to evaluate measurement precision of selection tests at the French pilot training. Int J Aviat Psych. 2012 Jan;22(1):18-29. doi: http://dx.doi.org/10.1080/10508414.2012.635123. PMID: 2012-01853-002. Exclusion: Pilot selection or training
177. Walters LC, Miller MR, Ree MJ. Structured interviews for pilot selection: No incremental validity. Int J Aviat Psych. 1993;3(1):25-38. doi: http://dx.doi.org/10.1207/s15327108ijap0301_2. PMID: 1993-32000-001. Exclusion: Pilot selection or training
178. Westerman R, Darby DG, Maruff P, et al. Computer-assisted cognitive function assessment of pilots. Australian Defence Force Health. 2001;2:29-36. Exclusion: Background
179. Wheeler JL, Ree MJ. The role of general and specific psychomotor tracking ability in validity. International Journal of Selection and Assessment. 1997 Apr;5(2):128-36. doi: http://dx.doi.org/10.1111/1468-2389.00052. PMID: 2011-05901-004. Exclusion: Pilot selection or training
180. Wilkening R. Relationship of CogScreen-AE to flight simulator performance and pilot age. Aviat Space Environ Med. 2000 Nov;71(11):1166. PMID: 11086677. Exclusion: Wrong Publication type
181. Wilson GF, Russell CA. Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Hum Factors. 2003 winter;45(4):635-43. PMID: 15055460. Exclusion: No included assessment
182. Woolsey SD. Relationship of CogScreen-AE to flight simulator performance and pilot age. Aviat Space Environ Med. 2001 Aug;72(8):771-2. PMID: 11506243. Exclusion: Wrong Publication type
183. Yakimovich NV, Strongin GL, Govorushenko VV, et al. CogScreen Performance of Russian Military Aviators 6 to 18 Months Following Closed Head Injury. 63rd Annual Scientific Meeting of the Aerospace Medical Association; 1995 Miami Beach, Florida, USA. Aviation, Space, and Environmental Medicine; May. Exclusion: Conference Abstract not enough data

184. Yesavage JA, Dolhert N, Taylor JL. Flight simulator performance of younger and older aircraft pilots: effects of age and alcohol. Journal of the American Geriatrics Society. 1994;42(6):577-82. Exclusion: Condition not included
185. Yesavage JA, Mumenthaler MS, Taylor JL, et al. Donepezil and flight simulator performance: Effects on retention of complex skills. Neurology. 2002;59:123-5. Exclusion: No included comparison
186. Zakay D, Shub J. Concurrent duration production as a workload measure. Ergonomics. 1998 Aug;41(8):1115-28. PMID: 9715671. Exclusion: No included comparison

Appendix F. Description of Included Studies

Table F-1: Distribution of Key Characteristics of Included Studies

	# of studies	% of studies
Decade of publication*		
1980-89	2	7%
1990-99	14	50%
2000-09	3	11%
2010-18	9	32%
Country		
US	21	75%
France	2	7%
Canada	1	3%
China	1	3%
Russia	1	3%
Singapore	1	3%
Mixed	1	3%
Type of Pilots		
Commercial air carrier/transport	3	11%
Commercial other	1	3%
General Aviation	4	14%
Military	10	36%
Mixed	3	11%
Astronaut	1	4%
Not reported	6	22%
Sample size+		
1-30	8	28%
31-50	2	7%
51-100	8	29%
101+	10	36%

*based on the date of first publication identified for a study; + based on the largest sample size used in any identified analysis

Table F-2: Key Characteristics of each included study

Study Author, Year	Number of Appraisal Criteria Met	Country	Type of Pilot	Total Sample Size	Sample population details
1.Barron, 2016	2	US	MIL	3,470	
2.Basner, 2015	1	US	Astronaut	19	8 astronauts, 11 mission controllers
3.Bell, 1997	2	US	MIL	40	
Waag 1994	1	US	MIL	205	
4.Carretta, 1996	3	US	MIL	117	
5.Causse, 2010	2	France	GA	24	
Causse, 2011a	3	France	GA	24	
6.Causse, 2011b	4	France	GA	32	
7.DeVoll, 2013	2	US	NR	98	
8.Doane, 2003	1	US	GA, MIL	77	
9.Endsley, 1990	1	US	MIL	25	Former military pilots, current status not reported
Endsley, 1994	0	US	MIL	25	Former military pilots, current status not reported
10.Griffin, 1987	2	US	MIL	22	
11.Hyland, 1994	1	US	COM	40	
12.Kay, 1995 (Clinical Studies II "Phase C Clinical Data") Sample 1 Sample 2	3	US	NR	100 145	60 pilots, 40 patients mixed sample of pilots and non-pilots (all clinical patients)
13.Kay 1995 (Clinical Studies I "FAA Phase B Study")/Kay 1991	4	US	NR	123	41 pilots, 42 healthy non-pilots, 40 patients
14.Kay, 1995 (US-Russia normative study)/Kay, 1993	3	Russia, US	COM	787	203 Russian, 584 US
15.Kay, 1995 (Clinical Studies III "Actual flight errors") / Yakimovich, 1994	4	Russia	COM	75	
16.Kennedy 2010	3	US	GA	72	
17.O'Donnell, 1992 (Le Roux, 1988)	4	US	NR	121	41 pilots, 40 healthy non-pilots, 40 patients *Subset for version 2.0* n=20 (5 pilots, 5 healthy non-pilots, 10 patients)
18.McGuire, 2014	2	US	MIL	170	
McGuire, 2016	3	US	MIL	216	
19.Moore, 1996	0	US	MIL	24	
20.Morrow, 2003	3	US	COM	187	91 pilots, 96 healthy non-pilots
21.Shull, 1990	1	US	MIL	66	
22.Stokes, 1991	4	US	NR	116	54 pilots, 62 patients
23.Stokes, 1992	3	US	GA, COM, MIL	25	

Study Author, Year	Number of Appraisal Criteria Met	Country	Type of Pilot	Total Sample Size	Sample population details
24.Sulistyawati, 2011	2	Singapore	MIL	16	
25.Taylor, 2000	4	US	GA, COM[1]	100	
Taylor, 2005	3	US	GA, COM[1]	97	
Taylor, 2007	3	US	GA, COM[1]	118	
Yesavage, 2011	3	US	GA, COM[1]	276	
Kennedy, 2013	3	US	GA, COM[1]	236	
Kennedy, 2015	3	US	GA, COM[1]	263	
26.Tolton, 2014	4	Canada	GA	54	
Van Benthem, 2016	3	Canada	GA	54	
27.Venturino, 1990	0	US	MIL	16	
28.Zhang 1997	3	China	NR	76	all participants are pilots; 63 healthy, 13 hospitalized

GA = general aviation; COM = commercial; MIL = military; n = number; NR = not reported; US = United States;
[1] Commercial limited to air transport
Note: The first article for each study is numbered. Rows that start with an author name instead of a number indicate that the article reports on the same study as the numbered row above.

Appendix G. Neuropsychological Tests

Simplified Assessment Name	in Battery (when applicable)	# of Pubs	Citations	Publications
Attention				
2-back Test		3	5-7	Causse, 2010; Causse, 2011a Causse, 2011b
Attention sharing	Situation Awareness Attribute Battery	1	10	Endsley, 1994
Attention/mental control	MicroCog	1	18	McGuire, 2014
Auditory letter span test		1	20	Sulistyawati, 2011
Auditory Sequence Comparison	CogScreen-AE	2	12,13	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1) Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study")
Backward digit span	CogScreen-AE WAIS-R	3	12-14	Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study") Kay, 1995 (Yakimovich 1994/Clinical Studies III "Actual flight errors")
Color-Word Test (modified Stroop)	Neuropsychological Test Battery	1	16	O'Donnell, 1992
Continuous opposites		1	4	Carretta, 1996
Dichotic listening task: single- and multitask		1	11	Griffin, 1987
Difference score: Visual Sequence Comparison and Divided Attention Test	CogScreen-AE	2	12,13	Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study")
Divided Attention Test	CogScreen-AE	4	12-14,19	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1) Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study") Kay, 1995 (Yakimovich 1994/Clinical Studies III "Actual flight errors") Moore, 1996
Dual Task Test	CogScreen-AE	6	13,14,24,26-28	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1) Kay, 1995 (Yakimovich 1994/Clinical Studies III "Actual flight errors") Taylor, 2007; Yesavage, 2011 Tolton, 2014; Van Benthem, 2016
Dual task: flight simulator and math (dual and single conditions)		1	29	Zhang, 1997
Dynamic memory test (Continuous Performance Test)	Neuropsychological Test Battery	1	16	O'Donnell, 1992

Simplified Assessment Name	in Battery (when applicable)	# of Pubs	Citations	Publications
Internal timing	Situation Awareness Attribute Battery Basic Attributes Test (adapted from)	1	[10]	Endsley, 1994
Matching to Sample	CogScreen-AE	5	[12-14,19,27]	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1) Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study") Kay, 1995 (Yakimovich 1994/Clinical Studies III "Actual flight errors") Moore, 1996 Tolton, 2014
Number comparison test		1	[20]	Sulistyawati, 2011
Perceptual vigilance	Situation Awareness Attribute Battery JAMJET (adapted from)	1	[10]	Endsley, 1994
Processing speed composite		1	[30]	Morrow, 2003
Scheduling 2		1	[4]	Carretta, 1996
Sentence-span test		1	[30]	Morrow, 2003
Spatial processing	Neuropsychological Test Battery	1	[16]	O'Donnell, 1992
Spatial Stroop		3	[5-7]	Causse, 2010; Causse, 2011a Causse, 2011b
Symbol Digit Coding	CogScreen-AE	3	[12,13,26]	Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study") Yesavage, 2011
Symbol digit substitution test	Neuropsychological Test Battery WAIS	1	[16]	O'Donnell, 1992
Time sharing 2		1	[4]	Carretta, 1996
Timers 1 & 2	MicroCog (as reaction time domain)	1	[18]	McGuire, 2014
Trail Making Test	Neuropsychological Test Battery	1	[16]	O'Donnell, 1992
Visual monitoring	Neuropsychological Test Battery	1	[16]	O'Donnell, 1992
Visual number span test (modified)		1	[20]	Sulistyawati, 2011
Visual Sequence Comparison	CogScreen-AE	4	[12,13,19,27]	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1) Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study") Moore, 1996 Tolton, 2014

Simplified Assessment Name	in Battery (when applicable)	# of Pubs	Citations	Publications
Working memory capacity		1	[8]	Doane, 2003
Working memory span composite	WAIS-R CogScreen-AE	1	[25]	Taylor, 2005
XYZ assignment		1	[4]	Carretta, 1996
Executive Function				
Maze task	Situation Awareness Attribute Battery	1	[10]	Endsley, 1994
Maze tracking test	SPARTANS	1	[32]	Stokes, 1992
Shifting attention	CogScreen-AE	10	[13,14,19,21,22,24-28]	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1) Kay, 1995 (Yakimovich 1994/Clinical Studies III "Actual flight errors") Kennedy, 2013; Kennedy, 2015; Taylor, 2005; Taylor, 2007; Yesavage, 2011 Moore, 1996 Tolton, 2014; Van Benthem, 2016
Memory				
ATC Situation Recognition task	Flitescript	1	[32]	Stokes, 1992
Building memory		1	[20]	Sulistyawati, 2011
Immediate/Delayed Memory	Situation Awareness Attribute Battery Basic Attributes Test (adapted from)	1	[10]	Endsley, 1994
Long-term working memory		1	[8]	Doane, 2003
Memory domain	MicroCog	1	[18]	McGuire, 2014
Spatial memory test	SPARTANS	1	[32]	Stokes, 1992
Sternberg memory retrieval	Neuropsychological Test Battery	1	[16]	O'Donnell, 1992
Motor Performance				
Interval production test	Neuropsychological Test Battery	1	[16]	O'Donnell, 1992
Laser aiming task 2		1	[4]	Carretta, 1996
Psychomotor task: single- and multitask		1	[11]	Griffin, 1987
Target hitting test		3	[5-7]	Causse, 2010; Causse, 2011a Causse, 2011b
Unstable tracking test	Neuropsychological Test Battery	1	[16]	O'Donnell, 1992
Perception				

Simplified Assessment Name	in Battery (when applicable)	# of Pubs	Citations	Publications
Aerial Orientation Test	Situation Awareness Attribute Battery	1	[10]	Endsley, 1994
Block counting	AFOQT	1	[31]	Barron, 2016
Block design test	WAIS-R	1	[30]	Morrow, 2003
Cube comparison test	Situation Awareness Attribute Battery	2	[10,20]	Endsley, 1994 Sulistyawati, 2011
Dot estimation	Situation Awareness Attribute Battery Basic Attributes Test (adapted from)	1	[10]	Endsley, 1994
Encoding speed	Situation Awareness Attribute Battery Basic Attributes Test (adapted from)	1	[10]	Endsley, 1994
Form board test		1	[20]	Sulistyawati, 2011
Group Embedded Figures test	Situation Awareness Attribute Battery	1	[10]	Endsley, 1994
Hidden figure test		1	[20]	Sulistyawati, 2011
Hidden patterns recognition	SPARTANS	1	[32]	Stokes, 1992
Manikin	CogScreen-AE	3	[12,13,24]	Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study") Taylor, 2007
Mental rotation ability	SPARTANS	1	[32]	Stokes, 1992
Perceptual speed	Situation Awareness Attribute Battery Basic Attributes Test (adapted from)	1	[10]	Endsley, 1994
Revised Minnesota Form Board Test	Situation Awareness Attribute Battery	1	[10]	Endsley, 1994
Rotated hidden patterns	SPARTANS	1	[32]	Stokes, 1992
Reasoning				
Analytic test	Situation Awareness Attribute Battery Graduate Record Examination	1	[10]	Endsley, 1994
Arithmetic Test	Neuropsychological Test Battery Unified Tri-Services Cognitive Performance Assessment Battery (adapted from)	1	[16]	O'Donnell, 1992
Deductive reasoning test		2	[5,6]	Causse, 2010; Causse, 2011a
Figure classification		1	[20]	Sulistyawati, 2011
Following directions		1	[20]	Sulistyawati, 2011

Simplified Assessment Name	in Battery (when applicable)	# of Pubs	Citations	Publications
Logical reasoning task nonsense syllogisms	SPARTANS	1	[15,16]	O'Donnell, 1992
Logical reasoning test	Neuropsychological Test Battery	1	[32]	Stokes, 1992
Math aptitude test		1	[20]	Sulistyawati, 2011
Math test	CogScreen-AE	3	[12,13,27]	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1) Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study") Tolton, 2014
Pathfinder	CogScreen-AE	6	[12,13,21,22,26,27]	Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1) Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study") Kennedy, 2013; Kennedy, 2015; Yesavage, 2011 Tolton, 2014
Rapid serial classification 4-square		1	[4]	Carretta, 1996
Raven's Advanced Progressive Matrices	Situation Awareness Attribute Battery	1	[10]	Endsley, 1994
Reasoning test		1	[7]	Causse, 2011b
Reasoning/calculation domain	MicroCog	1	[18]	McGuire, 2014
Verbal analogies		1	[31]	Barron, 2016
Verbal thinking test	Neuropsychological Test Battery	1	[16]	O'Donnell, 1992
Wisconsin Card Sorting test		3	[5-7]	Causse, 2010; Causse, 2011a Causse, 2011b
Situational Awareness				
Situation Awareness		1	[42]	Venturino, 1990
Situation Awareness Global Assessment Technique		3	[9,10,20]	Endsley, 1990; Endsley, 1994 Sulistyawati, 2011
Situation Awareness Rating Scale		2	[40,41]	Bell, 1997; Waag, 1994
Combination				
Combination: psychomotor and dichotic listening task measures		1	[38]	Shull, 1990
Cognitive ability tests		1	[30]	Morrow, 2003
General cognitive ability composite		1	[4]	Carretta, 1996

Simplified Assessment Name	in Battery (when applicable)	# of Pubs	Citations	Publications
Model: 2-back test + deductive reasoning test + total flight experience		1	[5]	Causse, 2010
Model: PASAT + Trail Making Test + Symbol Digit Modalities Test		2	[12,13]	Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study")
Psychomotor composite		1	[4]	Carretta, 1996
Speed of processing composite		3	[24-26]	Taylor, 2005; Taylor, 2007; Yesavage, 2011
Battery				
AMA mini-mental test		1	[36]	Stokes, 1991
Cognition		1	[33]	Basner, 2015
CogScreen-AE		5	[12,13,34,35,37]	DeVoll, 2013 Hyland, 1994 Kay, 1995 Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study") Kay, 1995 (Kay 1993/US-Russia normative study)
FFM mini-mental test		1	[36]	Stokes, 1991
Flitescript		1	[34]	Hyland, 1994
Illinois Screening Test version 1		1	[36]	Stokes, 1991
Illinois Screening Test version 2		1	[36]	Stokes, 1991
MicroCog		1	[18]	McGuire, 2014
Multidimensional Aptitude Battery–II		1	[18]	McGuire, 2014
Neuropsychological Test Battery		1	[16]	O'Donnell, 1992
SPARTANS		1	[36]	Stokes, 1991
WOMBAT		1	[34]	Hyland, 1994
Battery subset				
Air Force Officer Qualifying Test: subset	AFOQT	1	[31]	Barron, 2016

Simplified Assessment Name	in Battery (when applicable)	# of Pubs	Citations	Publications
CogScreen: subset	CogScreen-AE	11	12-14,19,21-23,27,28,37,39	DeVoll, 2013 Kay, 1995 (Clinical Studies II "Phase C Clinical Data" sample 1) Kay, 1995 (Kay 1991/Clinical Studies I "FAA Phase B Study") Kay, 1995 (Yakimovich 1994/Clinical Studies III "Actual flight errors") Kennedy, 2010 Kennedy, 2013; Kennedy, 2015; Taylor, 2000 Moore, 1996 Tolton, 2014; Van Benthem, 2016
MicroCog: subset	MicroCog	2	17,18	McGuire, 2014; McGuire, 2016

AE = Aeromedical Edition; AFOQT = Air Force Officer Qualifying Test; AMA = American Medical Association; ATC = air traffic control; ETS = Educational Testing Service; PASAT = Paced Auditory Serial Addition Test; SPARTANS = Simple Portable Aviation-Relevant Task-battery and Answer-scoring System; WAIS-R = Wechsler Adult Intelligence Scale-Revised

Appendix H. Glossary of Test Descriptions

Test Name	Domain	Description/Components
2-back test	Attention	Assesses working memory, specifically maintenance and updating; viewed continuous stream of stimuli and determine if it matched the shape of the stimulus 2-back in the sequence. Scored as percentage of correct responses.
Aerial orientation task	Perception	Measures a subject's ability to mentally rotate a two-dimensional aircraft outline. The subject's task was to select from five aircraft outlines presented at various rotations the one that showed the same side as a presented aircraft. The test consists of 30 items. Time to complete the test and accuracy was recorded.
AMA mini-mental test	Battery	Shortened version of FFM mini-mental test; includes registration, recall, attention, and calculation, plus 2 short-term memory tasks involving recall of digits and a spatial figure
Analytic Test	Reasoning	Designed to measure a subject's ability to understand a given structure of arbitrary relations among the presented items and to deduce new information from the relations given. For each question, there are five choices from which subjects are to select the correct response. Scored as number of correct responses. Subtest of the Graduate Record Examination; 25 questions with a time limit of 30 minutes.
Arithmetic reasoning	Reasoning	Provides a measure of the ability to understand arithmetic relations expressed as word problems. Test in the pilot composite score of AFOQT.
Arithmetic Test	Memory	Simple test of ability to carry out several addition and subtraction functions rapidly. Adapted from the Unified Tri-Services Cognition Performance Assessment Battery (UTCPAB).
ATC Situation Recognition task	Attention	Building a mental picture of a situation from air traffic control calls and selecting appropriate diagrams of the scenario.
Attention sharing	Attention	Computerized test consisting of a two-dimensional tracking task coupled with a digit cancellation task; subjects given 10 practice trials followed by three subtests of 1 minute each. A tracking task is presented during all three subtests based on a random-order, sinusoidal, rose-petal forcing function. Task difficulty ranged in value from 1 (easiest) to 10 (hardest). To keep the tracking task at the same level of perceived difficulty for all subjects across the testing period, the program automatically increased or decreased tracking difficulty to keep tracking errors at a constant prescribed level. In the digit cancellation task, a digit appeared at a random interval of between 5 and 15 sec. If subjects did not respond to the digit within 4 sec after its presentation, the tracking circle disappeared, forcing the subjects to cancel the digit in order to resume tracking. The tracking task was presented during all three subtests. In the first subtest, subjects canceled one of two digits (1 or 2) on the screen by pressing the corresponding key on a keyboard. In the second subtest, subjects canceled one of eight digits (1 to 8). In the third subtest, subjects performed only the tracking task. Scored as response time to cancel the digit, distance error for the tracking task, and average level of tracking difficulty.

Test Name	Domain	Description/Components
Attention/mental control domain	Attention	Subtests: Numbers forward, Numbers backward, Alphabet, Wordlist 1, Wordlist 2 Numbers forward & Numbers backward: "The two digit span subtests on MicroCog both present strings of digits on screen, one at a time, in increasing spans of 5–9 forward digits and 4–9 backward digits. Immediately after the string is presented, subjects must type the corresponding digits in their original (or reverse) order on the numeric keypad. If the subject responds correctly the following span is increased by one; if they answer incorrectly, the span is reduced by one." "Wordlist 1 requires that subjects press the Enter key whenever a word appears that belongs to a category specified in each of four trials. Two categories are phonemic, the other two semantic. Wordlist 2 uses the typical word list recognition format and presents the 16 words from Wordlist 1 imbedded with 20 others. Subjects respond whenever one of the words from the first list appears on the screen."
Auditory letter span test	Attention	Letters in series of varying length are read at a speed of one per second. Examinees instructed to write down the letters in the exact order in which they were called out. The examinees must not start writing until the series has been completed. Scored as number of series correct. From the 1976 edition of the Education Testing Service (ETS) Kit of Factor-Referenced Cognitive Tests; 24 series, about 10 minutes
Auditory Sequence Comparison	Attention	Comparison of two series of tone sequences.
Aviation information	NA[a1]	Evaluates knowledge of general aviation terms, concepts, and principles. Test in the pilot composite score of AFOQT.
Backward digit span (subtest of WAIS-R)	NA [a2]	Between three and six digits are displayed sequentially on the computer monitor, and the participant's task is to reproduce the sequences in the reverse order. The score is the percent accuracy for up to eight trials.
Backward digit span (subtest of CogScreen)	Attention	Sequential visual presentation of three to six digits. The subject's task is to reproduce the sequence in reverse order.
Block counting	Perception	Assesses spatial ability through the analysis of three-dimensional representations of a set of blocks. Test in the Combat Systems Officer Composite of the AFOQT.
Block design test	Perception	Test in the WAIS-R to measure spatial ability. Described as "a commonly used measure of visualization that accounts for age differences in performance of spatial memory and problem-solving tasks".
Building memory	Memory	The subject is asked to indicate the location of a number of buildings (12 items) seen on a previously studied map; provided 4 minutes for memorizing. From the 1976 edition of the Education Testing Service (ETS) Kit of Factor-Referenced Cognitive Tests; test time 4 minutes

Test Name	Domain	Description/Components
Cognition	Battery	Computerized cognitive test battery designed for astronauts, composed of 10 tests covering cognitive domains of emotion processing, spatial orientation, and risk decision making: motor praxis task, visual object learning test, fractal 2-back, abstract matching, line orientation test, emotion recognition task, matrix reasoning test, digit-symbol substitution task, balloon analog risk test, and psychomotor vigilance test. Tests incorporated from the Penn Computerized Neurocognitive Battery (CNB): motor praxis task, emotion recognition task, and matrix reasoning test; tests adapted from the Penn CNB: line orientation test, fractal 2-back; test adapted from the WAIS-III: digit-symbol substitution task. Tests scored based on accuracy and speed, with weighting differing by subtest; standardized score range 0 to 1000 for each subtest with final score as sum of all subtests.
Cognitive ability tests	Combination	Cognitive ability tests entered together in model for analysis: sentence-span test, processing speed composite, and block design test.
CogScreen-AE	Battery	A computer-administered and scored cognitive-screening instrument designed to rapidly assess deficits or changes in attention, immediate- and short-term memory, visual-perceptual functions, sequencing functions, logical problem solving, calculation skills, reaction time, simultaneous information processing abilities, and executive functions. The battery consists 13 subtests, each of which may be scored in multiple ways (called "measures"), such as speed, accuracy, and thruput (number of correct responses per minute), depending on the individual subtest.
Color-Word Test (modified Stroop)	Attention	Requires the subject to name the color in which a word is written even though the word may be the name of that color or of a different color.

Test Name	Domain	Description/Components
Combination: psychomotor and dichotic listening task measures	Combination	Combination: psychomotor and dichotic listening task measures (total 15 measures/scoring methods); both tasks performed in single and multitask conditions (multitask = simultaneous performance of both tasks) - Psychomotor task (PMT): Required subjects to maintain first one, then two, and finally three, randomly displaced cursors on fixed targets on the CRT by manipulating joysticks and foot pedals. Subjects manipulated one joystick, located at the front seat edge, with their right hand to control a cursor that moved within the upper two-thirds of the screen just right of center in a backwards (reversed) manner. Locally produced rudder pedals were used to control a cursor that moved horizontally across the bottom of the screen. Pushing the left pedal moved this is cursor to the right while pushing the right pedal moved it to the left. Another joystick (throttle), located on the left seat edge, was manipulated by the subject's left hand to move a cursor vertically on the left side of the screen. The subject pulled this throttle back to move this cursor down and vice versa. Psychomotor task test scores were the accumulated total of absolute errors from an ideal target position. - Dichotic listening task (DLT): A series of letter/digit string sets presented to subjects aurally over binaural headphones via two voice synthesizers. Subjects were told which ear to attend to for each trial. Part I was a series of 16 pairs of letters and/or numbers; Part II was a series of 6 more pairs. Subjects were to indicate the digits (0-9) presented to the designated ear in the order of their occurrence. Subjects responded with their left hand using a separate keypad placed immediately in front and slightly left of center. The test was preceded by six aural practice trials, which provided immediate performance feedback by visually indicating the letters and digits presented and the subjects' keypad responses. Subjects also completed three multiple-choice questions before beginning the actual test to ensure that they understood the concept of the DLT. The DLT performance measure was the number of incorrect responses during 12 trials in which a total of 108 correct responses were possible. - Multitask condition: Subjects performed both the DLT and PMT simultaneously (a 12-trial DLT and a 4.5-min PMT). During the first multitask condition, subjects performed the DLT and the stick-only PMT. During the next two multitask conditions, subjects performed the DLT and the stick-and-rudder PMT using their right hand and feet to control the central joystick and the rudder pedals, and their left hand to make keypad responses to the DLT input. During the final multitask condition, subjects performed the DLT and the stick-rudder-and-throttle PMT. In this most elaborate combination, subjects used their right hand and both feet to control the central joystick and the rudder pedals as before but, in addition, used their left hand to control the throttle joystick and voiced their DLT responses using a microphone attached to the headphones. Performance measures for the PMT and DLT in these multitask conditions were identical to those of the single tasks with PMT errors being recorded for the final 4 min of that test.

Test Name	Domain	Description/Components
Computation span	NA [a2]	The participant sees a sequence of up to nine arithmetic problems (such as 6 + 2 =?) and three response alternatives. As the participant answers the problems in a sequence, he or she is asked to remember the last digit of each problem in that sequence (2, in this example). The first trial had one arithmetic problem and one memory item. If the participant correctly recalls all the memory items on three consecutive trials, then the number of problems (and memory items) is increased by one. Computation span is scored as the largest number of digits the participant could correctly recall on at least two out of three test trials.
Continuous Opposites	Attention	Test of verbal working memory. Participants are required to remember the last three words (or their opposites) in a list presented one word at a time. If a word appears in the color red, the participant is required to remember its opposite.
Cube comparison test	Perception	Measures ability at mental rotation in 3 dimensions. Subjects presented with 21 drawings of pairs of cubes (similar to children's blocks) that had a unique letter or number on each face of the cube. The task is to determine if the two drawings could represent the same block by mentally rotating the blocks so that they would have the same orientation.
Deductive reasoning test	Reasoning	The goal of the task is to solve syllogisms by choosing, among three suggested solutions, the one that allows concluding logically. Syllogisms are based on a logical argument in which one proposition (the conclusion) is inferred from a rule and another proposition (the premise). We used four existing forms of syllogisms: modus ponendo ponens, modus tollendo tollens, setting the consequent to true, and denying the antecedent. Each participant had to solve 24 randomly displayed syllogisms. The measurement was the percentage of correct responses.
Dichotic listening task: single- and multitask	NA[b]	See “Combination: psychomotor and dichotic listening task measures” - appears to be same (no description provided).
Digit copying task	NA [a3]	One task in a speed of processing test (the other being pattern comparison); number of items correctly completed transformed to z-score for each task, then z-scores averaged to give composite score of speed of processing.
Divided Attention Test	Attention	Subject monitors the vertical movements of a bar within a circle and returns the bar to the center position when its deviation from center exceeds an upper or lower boundary. The monitoring task is presented alone and in combination with the Visual Sequence comparison task.
Dot estimation	Perception	The test is computerized and consists of 50 test trials with no practice trials. During the test, subjects are shown two equal sized square boxes on a computer screen. The boxes contained a number of white dots, with one of the boxes containing one more dot than the other. The subject's task was to indicate as quickly as possible which box contains more dots, using designated keys on the keyboard. Accuracy and response time are recorded for each trial.

Test Name	Domain	Description/Components
Dual task: flight simulator and math	Attention	Task performed without flight (baseline) or during flight (as secondary task). Computerized operation task and value comparisons of numeric stimuli. Subject presented with arithmetic problems of 3-digit addition and subtraction then must respond by indicating whether the obtained sum was greater or less than a prespecified value of five using two button keys on the throttle by the left hand. The arithmetic problems were randomly generated using only the numbers 1-9. The problem changed either after subject responded or after a 4-second time limit. Scores include reaction time, error, lost rate (no response in time limit), and Information Processing Speed calculated as IPS = {[1 - E/(N + E)]*LOG2 K)/T, where N: total times of each trial; E: error times on trial; T: averaged correct reaction time on each trial; and K: times of probe presented randomly.
Dual Task Test	Attention	Consists of two tasks, each of which is performed alone and then together as a simultaneous task. One task is a visual-motor tracking test. The second task is a continuous memory task, involving recall of the previously presented number.
Dynamic memory test (Continuous Performance Test)	Attention	Requires the subject to note the bottom number of a fraction. When a new fraction appears, the subject must respond by saying whether the top number is the same as the previous bottom number. However, the new bottom number must first be noted because as soon as a response is given the original fraction is replaced by a new one.
Encoding speed	Perception	The test is computerized consisting of 3 subtests of 32 trials each and 10 practice sessions for each subtest. Subjects were presented with two pairs of letters and had to decide whether the pairs of letters were the same or different. The pairs remained on the computer screen until the subjects responded. Each subtest used a different rule for similarity: physical identity, letters in both pairs must be identical in letter and in case (AA and AA), name identity, both pairs of letters must be composed of the same letter regardless of case (AA and Aa) or categorical identity, and letter pairs need to be either all vowels or all consonants (Ai and Ea). Accuracy and response time are recorded.
FFM mini-mental test	Battery	assesses 5 areas of cognitive functioning: orientation in time and space, "registration" (naming 3 objects), attention and calculation (decrementing a value by sevens aloud), language skills (following spoken instructions, repeating and generating sentences), and recall (of 3 objects)
Figure Classification	Reasoning	Each item presents 2 or 3 groups each containing 3 geometrical figures that are alike in accordance with some rule. The second row of each item contains 8 test figures. The task is to discover the rules and assign each test figure to one of the groups. Scored as number identified correctly minus a fraction of those incorrect. From the 1976 edition of the Education Testing Service (ETS) Kit of Factor-Referenced Cognitive Tests
Flitescript	Battery	Index pilots' representations of situational knowledge in long term memory There are two versions of the Flitescript test, a recall version and a recognition version. The recall version of the test involves reconstructing both randomized and coherent air traffic control (ATC) radio call sequences from memory. The recognition version requires listening to an ATC communication sequence and selecting the correct graphic depiction of the situation represented by the ATC communications from a set of alternatives.

Test Name	Domain	Description/Components
Following Directions	Reasoning	The subject is asked to determine the point in a matrix of letters that would be reached by following a complex set of directions. Scored as number of letters marked correctly minus a fraction of those incorrect. From the 1976 edition of the Education Testing Service (ETS) Kit of Factor-Referenced Cognitive Tests; 10 items, 7 minutes
Form board test	Perception	Each items presents 5 shaded drawings of pieces, some or all of which can be put together to form a figure presented in outline form. The task is to indicate which of the pieces, when fitted together would form the outline. Scored as number correct minus number incorrect. From the 1976 edition of the Education Testing Service (ETS) Kit of Factor-Referenced Cognitive Tests; 24 items, 8 minutes.
Forward digit span	NA [a2]	Requires subjects to repeat a series of digits of increasing length.
General cognitive ability composite	Combination	General cognitive ability composite = unit-weighted sum of 4 tests (continuous opposites; rapid serial classification; scheduling 2; XYZ assignment)
General cognitive functioning domain	Battery Subset	MicroCog domain. Combines the information processing speed and accuracy scores, giving equal weight to each.
General cognitive proficiency domain	Battery Subset	MicroCog domain. Combines speed and accuracy but is based on proficiency scores and thus gives preferential weight to accuracy over speed.
Group Embedded Figures test	Perception	Each problem in the test consists of a complex geometric pattern that contains one of eight simple geometric figures that are presented to the subject. Subjects were required to trace with a pencil the simple geometric figure that is embedded in each complex pattern. Three minutes are provided for a practice section with seven problems, and 5 minutes for each of two test sections with eight problems. Accuracy and time to complete each section are recorded.
Hidden figure test	Perception	The task is to decide which of 5 geometrical figures is embedded in a complex pattern. Scored as number marked correctly minus a fraction of the number marked incorrectly. From the 1976 edition of the Education Testing Service (ETS) Kit of Factor-Referenced Cognitive Tests; "adaption of the Gottschaldt Figures type test", "the difficulty level of this test is high"; 16 items, 12 minutes.
Hidden patterns recognition	Perception	Assesses flexibility of closure and factor loads with spatial ability. Subjects must detect an abstract line drawing embedded within a more complex pattern of lines.
Illinois Screening Test version 1	Battery	Assesses skills for performance in complex dynamic environments such as flying; two neuropsychological tests for each of 6 skills in each battery (version 1 or 2), skills are: perceptual-motor ability, spatial ability, working memory capacity, attentional performance (visual scanning), processing flexibility, and planning/sequence ability
Illinois Screening Test version 2	Battery	see description for version 1

Test Name	Domain	Description/Components
Immediate/delayed memory	Memory	The test is computerized, and consists of two subtests, during which a series of one-digit numbers are flashed on a computer screen for .05 seconds. The interstimulus interval was 2 seconds for half of each subtest (immediate) and 5 seconds for the remainder (delayed). In the first subtest, subjects were asked to respond via the keyboard with the number that appeared immediately prior to the number on the screen. In the second subtest, subjects were asked to respond with the number that appeared two numbers prior to the number displayed on the screen. For both subtests, 10 practice trials and 50 test trials were conducted. Accuracy and response time for each subtest was recorded.
Information processing accuracy domain	Battery Subset	MicroCog domain. Measures the accuracy of performance with no regard given to speed.
Information processing speed domain	Battery Subset	MicroCog domain. Measures the time it takes an individual to complete simple and complex mental tasks.
Instrument comprehension	NA [a5]	Assesses the ability to determine the attitude of an aircraft from illustrations of flight instruments. Test in the pilot composite score of AFOQT.
Internal timing	Attention	Subjects viewed three labeled points (A, B, and C) placed in a straight line on a computer screen. When subjects depress the space bar on the keyboard, a target begins moving from A at a constant velocity and is blanked from the screen as it passes B. The subject's task was to determine when in time the target would reach C and to press the space bar at that time in response. Adapted from Basic Attributes Test.
Interval production test	Motor Performance	Requires the subject to tap at a regular rate two to three per second for three minutes.
Laser aiming task 2	Motor Performance	Tests coordination and aiming. Participants are instructed to imagine they are shooting from an aircraft at the bottom of the screen. Participants must match the apparent altitude (size) of the target and the laser un to get the laser beam on target.
Letter comparison	NA [a3]	Part of the processing speed composite (other task is pattern comparison). Participant decides whether pairs of letter are the same or different, as fast as possible.
Logical Reasoning Task Nonsense Syllogisms	Reasoning	Subject is presented with a series of syllogisms and must decide in each case whether the conclusion is valid or invalid. Adapted from the ETS test.
Logical Reasoning Test	Reasoning	A series of symbols are presented, along with a verbal description of the logical relationships between them. The subject must determine whether the logical relations described are true or not with respect to the presented symbols.
Long-term working memory	Memory	Subjects simultaneously view 2 cockpits for 40 seconds. One cockpit appears on the top half of the screen, with the second cockpit displayed directly below the first. After 40 seconds lapses the computer presents a number and asks the subject to count backwards aloud by threes for 30 seconds starting from the presented number. Then the computer prompts the subject to recall situation specific values displayed in either the top or bottom cockpit. Subject use a sheet of paper containing the 7 instruments with no values and a pen to fill in the situation specific values for each instrument. In six trials the cockpits are related. Three trials consist of two unrelated cockpits.

Test Name	Domain	Description/Components
Manikin	Perception	Subject identifies the hand in which a rotated human figure is holding a flag.
Matching to sample	Attention	Following a brief presentation of a four-by-four pattern of colored squares. The subject identities the matching pattern from two choices.
Math knowledge	NA [a5]	Measures the ability to use mathematical terms, formulas, and relations. Test in the pilot composite score of AFOQT.
Math test	Reasoning	Traditional math word problems with multiple choice answer format.
Mathematics aptitude test	Reasoning	Test consists of 5-choice word problems requiring arithmetic or very simple algebraic concepts only. Scored as number correct minus a fraction of incorrect. From the 1976 edition of the Education Testing Service (ETS) Kit of Factor-Referenced Cognitive Tests; 15 items, 10 minutes. Overlaps with Arithmetic Aptitude Test in difficulty.
Maze task	Executive Function	Task to measure subjects' abilities at spatial orientation on a fixed map. The task consists of four three-dimensional mazes. A practice maze was given to each subject to familiarize him with the task. Total time to reach the endpoint of each maze successfully was recorded.
Maze tracking test	Executive Function	Line mazes of increasing complexity. Each maze must be cognitively traced as rapidly as possible to decide whether or not there is an unbroken path from beginning to end. Adapted from Educational Testing Service kit of Factor-Referenced Cognitive Tests
Memory domain	Memory	Domain on MicroCog. Subtests: Story 1, Story 2, Address "Address 1 asks subjects to simply read and remember a fictional name and address. It scores only the time it takes subjects to indicate they have learned the material. Address 2 later asks multiple-choice questions about the same information. MicroCog presents two short stories (only one on the short version), separated by several other subtests. Subjects read each story and answer multiple-choice questions immediately and again after a delay. Because both memory subtests are multiple choice tests they measure recognition, not recall."
Mental rotation ability	Perception	See description for "rotated hidden patterns", no description provided in publication, but appears to be the same.
MicroCog	Battery	Computer-administered test battery containing 18 subtests that contribute to summary scores for nine interrelated cognitive areas (domains): Attention/ Mental Control, Memory, Spatial Processing, Reasoning/Calculation, Reaction Time, Information Processing Accuracy, Information Processing Speed, Cognitive Functioning, and Cognitive Proficiency.
Multidimensional Aptitude Battery–II	Battery	Broad-based test of intellectual and cognitive ability. The test yields a full-scale intelligence quotient 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.

Test Name	Domain	Description/Components
Neuropsychological Test Battery, version 1.0	Battery	Test battery using step approach to testing (3 steps); validation experiment in test battery development with candidate tests and level designation based on preliminary study. Level 1 tests: trails A; trails B; symbol digit; color word; unstable tracking. Level 2 tests: continuous performance; verbal thinking; arithmetic; interval production. Level 3 tests: spatial thinking; memory test; visual monitor; logical reasoning; Zung depression; manifest anxiety; Shipley Scale.
Neuropsychological Test Battery, version 1.1	Battery	Test battery using step approach to testing (3 steps); revised battery (second-generation breadboard) based on validation experiment (version 1.0) in test battery development. Level 1 tests: trails A; trails B; symbol digit; tracking. Level 2 tests: logical reasoning (% correct); dynamic memory (S.D.); arithmetic (attempts). Level 3 tests: memory (slope); Zung depression; manifest anxiety; dynamic memory (reaction time)
Neuropsychological Test Battery, version 2	Battery	Test battery using step approach to testing (3 steps); version 1.1 tests all administered by computer. Level 1 tests: trails A; trails B; symbol digit; tracking. Level 2 tests: logical reasoning (% correct); dynamic memory (S.D.); arithmetic (attempts). Level 3 tests: memory (slope); Zung depression; manifest anxiety; dynamic memory (reaction time).
Number comparison test	Attention	The subject inspects pairs of multi-digit numbers and indicates whether the two numbers in each pair are the same or different. Scored as number correct minus number incorrect. From the 1976 edition of the Education Testing Service (ETS) Kit of Factor-Referenced Cognitive Tests; 48 items, 1.5 minutes,
Pathfinder	Reasoning	Adapted from trail making test. After viewing a number or letter displayed in the center of the screen, the respondent's task is to select one of the four quadrants containing the next character in a previously specified sequence. Three of the four characters are updated following each response. Three sequencing rules: number, letter, or combined (alternating series of numbers and letters. Performance measured as: speed (median response time for correct responses); response accuracy (% correct responses); thruput (# correct responses per minute); coordination measures indicating proximity to center of target numbers and letters (mean deviation from center of target).
Pattern comparison task	NA [a2]	Participant makes same-different decisions about pairs of patterns made of connected line segments as quickly as possible. As task in a speed of processing test: number of items correctly completed transformed to z-score for each task (other task being digit copying), then z-scores averaged to give composite score of speed of processing. As task in the processing speed composite: mean processing speed score created from scores on pattern comparison and letter comparison tasks.
Perceptual speed	Perception	Computerized test consisting of a practice session of 10 trials and 5 subtests of 16 trials each, with stimulus presentation times of 500, 400, 300, 200, and 100 ms, respectively. A three- to seven-digit number was presented on the computer screen for the prescribed stimulus time. After a 500-ms delay, a second number was presented. Subjects had to decide as quickly as possible whether the two numbers were the same or different (due to transposed digits) by pressing designated keys on the keyboard in response. Scored for accuracy and reaction time.

Test Name	Domain	Description/Components
Perceptual vigilance	Attention	Task designed to measure monitoring and instrument-scanning abilities. Subjects were shown a computer screen with 25 rows of 80 red dots on a black background. Subjects were instructed to scan the screen thoroughly for a change in one of the dots from red to magenta that was just above visual threshold. The subject signaled when they noticed the change by pressing any key on the keyboard. 10 trials with stimulus-onset intervals ranging from 1 to 15 sec were administered. Elapsed time from the color change to the subject's response was recorded.
Pilot Composite Score	Battery Subset	AFOQT composite score generated from 5 tests: arithmetic reasoning, math knowledge, instrument comprehension, table reading, and aviation information.
Processing speed	NA[b]	see description for "Speed of Processing"
Processing speed composite	Attention	Letter comparison and Pattern comparison tasks. Scored as a mean processing speed score of both tasks.
Psychomotor composite	Combination	Psychomotor composite = unit-weighted sum of 2 tests (laser aiming task 2; time sharing 2).
Psychomotor task: single- and multitask	Motor Performance	See description for the psychomotor task under “Combination: psychomotor and dichotic listening task measures"; no description provided in publication, but assumed to be the same psychomotor task.
Psychomotor Vigilance Test	NA [a4]	Records reaction times to visual stimuli that occur at random inter-stimulus intervals. Subjects are instructed to monitor a box on the screen, and hit the space bar once a millisecond counter appears in the box and starts incrementing. The reaction time will then be displayed for 1 second. Subjects are instructed to be as fast as possible without hitting the spacebar without a stimulus.
Rapid serial classification: 4-square	Reasoning	Participants are shown a 4-square (2x2) display in which a letter pattern can be drawn (C, X, or Z) between points. Participants must determine which letter is being drawn by following the pattern of dots as they are sequentially illuminated and extinguished.
Raven's Advanced Progressive Matrices	Reasoning	Tests ability to perform pattern recognition using nonverbal reasoning skills. Subjects are shown a pattern with a piece missing and instructed to select the missing piece from eight choices. 12 familiarization problems followed by 36 problems arranged in increasing order of difficulty. Scored as number of correct responses.
Reasoning test	NA[b]	see description for Deductive Reasoning Test

Test Name	Domain	Description/Components
Reasoning/calculation domain	Reasoning	Domain of the MicroCog. Subtests: Analogies, Object Match A & B, Math calculation Analogies: Presents a series of 11 relationships (x is to y as a is to …), which subjects must choose among the choices displayed. Object Match: Presents each of 12 frames of four figures. Each frame is presented twice. Subjects must choose which figure is different because of some dimension, such as color, size, or shape. Having selected a figure according to one dimension on the first presentation of each frame, subjects must use another dimension to select a different figure when that frame is repeated. Thus, each frame has two correct answers, which subjects can provide in either order. Object Match A consists of first trial of each frame, Object Match B the second. Note: "Object Match closely resembles the Category Test but differs from both it and the WCST by not informing subjects whether they are responding correctly. MicroCog provides feedback only when subjects attempt to repeat their first response in the same frame." Math calculation: eight arithmetic problems, which subjects solve without paper or pencil and enter the 1–4 digit answers on the number keyboard.
Revised Minnesota Form Board Test	Perception	Designed to measure ability to visualize and manipulate 2-dimensional geometric shapes into a whole design. The test consists of 64 two-dimensional diagrams, each comprised of a collection of pieces analogous to an unassembled jigsaw puzzle. For each diagram, the subject is to select from five possible answers the one that correctly represents the pieces put together as a whole figure. Scored as number of correct responses in 20 minutes.
Rotated hidden patterns	Perception	Subjects must detect an abstract line drawing embedded within a more complex pattern of lines, and the target figure may be rotated. Mirror images are considered non-targets.
Scheduling 2	Attention	Five horizontal logarithmic scales can be presented. A line beneath each scale increases at a unique, constant rate. Each line and scale appears on a separate screen which may be viewed by entering the scale number on the response keypad. Participants score points equal to the current value of the line displayed on the scale by pressing the ENABLE key. When the ENABLE key is pressed, the participant's total score is incremented by the value of the line, which is then reset to 0, where it will start increasing again. If the value of a line reaches the upper limit of the scale and the participant has not responded by pressing the ENABLE key, the value of the line will reset to 0 without the participant receiving any points.
Sentence span	NA [a2]	Participants answer questions about sentences, and their task is to remember the last word of each sentence in a sequence.
Sentence-span test	Attention	Contains a listening and reading component measuring "the ability to simultaneously store and manipulate information in memory". Scored as mean span score of listening and reading components.

Test Name	Domain	Description/Components
Shifting attention	Executive Function	Shifting Attention Test (SAT) begins by training the respondent to select from among four response boxes, each of which contains an arrow, according to one of three easily learned rules: The first rule requires the respondent to select a box based on the color of its border; the second rule requires the respondent to select a box based on the direction of the arrow; and the third rule requires the respondent to match the color of the arrow. After learning the three response conditions, the respondent begins the fourth condition, SAT Instruction. In this condition. The respondent is presented with an instruction identifying the active rule before the presentation of each subsequent stimulus. In the fifth condition. SAT Discovery. The respondent's task is to discover and then apply the active response rule which changes after a variable number of correct responses. The respondent uses trial-and-error to ascertain the active rule. Scored for individual “measures”: For each task of Arrow Direction, Arrow Color, Instruction, and Discovery: scored for Accuracy, Speed, and Thruput For discovery task only: scored for Rule Shifts Completed; Failures to Maintain Set; Nonconceptual Responses; Perseverative Errors
Situation Awareness (observer and self-report)	Situational Awareness	Subjective ratings collected at the time of simulation, and obtained from 3 sources (pilot self-rating, flight lead, and trained observer). These ratings were based on planned tactics, executed tactics, an assessment of the flight's response to tactical situations, and the participants' awareness of ongoing tactical events, independent of the final engagement outcome.
Situation Awareness Attribute Test Battery	Situational Awareness	18 tests covering 5 attribute areas. Attribute areas: spatial; attention; memory; perception; cognitive
Situation Awareness Global Assessment Technique (SAGAT)	Situational Awareness	Uses a freeze probe technique. As a global measurement tool, SAGAT includes queries about all operator SA requirements, including level 1 (perception of data), level 2 (comprehension of meaning) and level 3 (projection of the near future) components. This includes a consideration of system functioning and status as well as relevant features of the external environment. By including queries across the full spectrum of an operator’s SA requirements, this approach minimizes the possible bias of attention, as subjects cannot prepare for the queries in advance.
Situational Awareness Rating Scale (SARS)	Situational Awareness	The self-report SARS and supervisory SARS required the respondents to rate either themselves or their subordinates on each of the 31 items. Scale anchors were "Acceptable" and “Outstanding"
SPARTANS automated battery	Battery	Designed to assess individual differences in pilots; 9 sub-tasks: dual task tracking, risk taking task, Sternberg task, maze tracing, hidden figures, visual scanning, schedule task, dual task decrement, visual number span
Spatial memory test	Memory	In this test the subject views an inspection set of nonsense figures. These are abstract amoeboid figures without geometrical or pictorial significance that would facilitate verbal recording. About twenty minutes after initial presentation subjects view 40 figures and decide for each one whether or not it had been a member of the inspection set viewed previously.

Test Name	Domain	Description/Components
Spatial processing	Attention	A four-bar histogram is presented. After 3 seconds it is removed and replaced (after a delay) with another histogram rotated either 90 or 270 degrees. The subject must decide whether the second histogram is the same as the first. Intact spatial memory is required, as well as ability to mentally manipulate spatial symbols.
Spatial processing domain (MicroCog)	Battery Subset	Domain on the Microcog comprised of Tic Tac (2 subtests), and Clocks "Clocks displays seven clock faces in turn, with hands but no numbers. Subjects choose the correct time from among the five choices given. The first few faces include hour markers, such as those on analog wristwatches; these are deleted on later presentations. The two Tic Tac subtests are based on a 3 × 3 grid. After a brief introduction, in which subjects are shown which numbers correspond to the grid positions, they are shown random frames of 3–5 squares, each frame appearing for 1s. MicroCog presents the squares simultaneously, not sequentially like block tapping. After each frame, subjects must designate the locations of the squares by typing the corresponding numbers on the keyboard to match the same sequence. Thus they must first learn to associate the 3 × 3 grid with numbers 1–9.
Spatial Stroop	Attention	Assesses inhibition efficiency with conflict between the meaning of a word naming a location and the location where the word is displayed. Stroop neutral meaning = motor answer given with appropriate hand according to the meaning. Stroop neutral position: response given according to location of string of XXXXX displayed at the left or right of the screen. Stroop meaning incompatible/compatible: response is given according to the meaning of the word, compatible or incompatible with its location at the screen. Scored as interference score = stroop meaning incompatible - (stroop neutral position * stroop neutral meaning) / (stroop neutal position + stroop neutral meaning)
Speed of processing	Combination	Comprised of 2 tasks: pattern comparison and digit copying; number of items correctly completed transformed to z-score for each task, then z-scores averaged to give composite score of speed of processing. In the Pattern Comparison task, participants were asked to make same–different decisions about pairs of patterns made of connected line segments. In the Digit Copying task, participants were asked to copy digits as rapidly as possible. For each task, participants were given two 30-s trials, and the number of correct responses was scored. The two scores were standardized and then averaged together to provide a composite measure of speed.
Speed of processing standardized composite	Battery Subset	Speed of processing standardized composite: standardized variables - pathfinder throughput and symbol digit throughput (both from CogScreen subtests: pathfinder, symbol digit coding).
Sternberg memory retrieval	Memory	Determining whether a "probe" letter of the alphabet is a member of a previously memorized target set using Sternberg paradigm.
Symbol Digit Coding	Attention	Substitution of digits for symbols using a key followed by testing of immediate and delayed recall of symbol-digit pairs.
Symbol digit substitution test	Attention	Requires the subject to substitute numbers for geometric symbols.
Table reading	NA [a5]	Measures the ability to quickly and accurately extract information from a table at a given set of X and Y coordinates. Test in the pilot composite score of AFOQT.

Test Name	Domain	Description/Components
Target hitting	Motor Performance	Participant instructed to click as fast as possible on each target. Test of basic psychomotor reaction time, measured by velocity index (velocity index = average ratio of base 10 log of distance in pixels between 2 targets divided by time in seconds to go from 1st to 2nd target)
Time sharing 2	Attention	Measures attention, reaction time, and rate control. First part of test is a compensatory tracking task, second part is an attention task, with simultaneous tracking and attention tasks. Part 1, compensatory tracking: "participants maneuver the right-hand control stick to keep a 'gunsight' centered on an airplane"; Part 2, attention: "Numbers appear one at a time in sequence at the lower part of the screen... Occasionally, a number will be missing from the sequence. Participants are required to type the missing number on the keypad." Part 3 combined: "during the final part of this test, participants simultaneously perform tracking and attention tasks"
Timers 1 & 2	Attention	Tests comprise the reaction time domain on MicroCog. Subjects must press the Enter key to respond successively to five aural tones, five on-screen images, and then five images proceeded briefly by a tone.
Trail Making Test A	Attention	Twenty five numbered circles are to be joined in sequence
Trail Making Test B	Attention	Twenty five numbered circles are numbered 1 to 13 and A to L and are to be joined in alternating sequence.
Unstable tracking test	Motor Performance	The subject must keep a computer generated "target" centered with a tracking knob or a joystick while the computer generates offsets for the target.
Verbal analogies	Reasoning	Measures the ability to reason and determine relationships between words. Test in the Verbal Composite score of the AFOQT.
Verbal Thinking Test	Reasoning	The subject has to classify two letters of the alphabet by each of two rules. One rule involves physical identity alone (whether both are the same letter in the same case). The other involves a semantic rule (whether they both are vowels or consonants).
Visual monitoring	Attention	Requires the subject to monitor four dials (similar to aircraft dials) to detect a randomly occurring bias in one of them.
Visual number span test (modified)	Attention	Items presented by having each digit printed on a large card, flipping over one card per second, or otherwise exposing one digit per second, for the examinees to see. Examinees instructed to write down the numbers in the exact order in which they were shown. The examinees must not start writing until the series has been completed. Scored as number of series correct. From the 1976 edition of the Education Testing Service (ETS) Kit of Factor-Referenced Cognitive Tests; 24 series; about 10 minutes.
Visual Sequence Comparison	Attention	Comparison of two simultaneously presented series of letters and numbers.
Wisconsin Card Sorting Test	Reasoning	Computerized: sort cards according to 3 different unknown categories (color, shape, number) with audio feedback as yes/no for correct response; target category automatically changed when participant successfully categorized 10 cards; task ended when 6 categories achieved (color, shape, number, color, shape, number) or deck of 128 cards used. Scored as total number of perseverative errors (≥2 unsuccessful sorting on the same category).

Test Name	Domain	Description/Components
WOMBAT	Battery	Measures the ability of the test participant to simultaneously perform several tasks and to determine changing priorities associated with task execution. This requires that the test participant judge the relative worth of a particular action at each moment and dynamically reorder task priorities. Requires that the test participant develop a strategy for dealing with constantly changing priorities. Provides a rigorous test of the pilot's ability to attend to varying sources of information and to shift priorities appropriately. Provides a measure of vigilance through a comparison of mid- and end-test segment scores.
Word knowledge	NA [a5]	Assesses verbal comprehension involving the ability to understand written language through the use of synonyms. Test in the Verbal Composite score of the AFOQT.
Working memory capacity	Attention	A combined analog measure of both verbal working memory and spatial working memory. Subjects view a series of altitude indicator displays positioned in different flight orientations. Upon presentation, subjects were asked to say aloud whether the aircraft was pitched up or down. Subjects are asked to remember the orientation of the horizon line displayed on the altitude indicator and the number below the altitude indicator. Subjects first view of two altitude indicators for 5 trials and progress through a series length of three and four, each containing five trials.
Working memory span composite	Attention	Working memory span composite is comprised of 5 tests: computational span, sentence span, forward digit span (from WAIS-R), backward digit span (from WAIS-R), and visual backward digit span accuracy (from CogScreen).
XYZ assignment: synthesis add and subtract	Attention	Participants are required to combine or delete simple line figures assigned to three letters (X, Y, and Z). Two figures are assigned to each letter in the form of an addition or subtraction equation. Participants must mentally combine or delete the lines of these figures and then memorize the combination. Information about one figure is sometimes needed to solve the equation for one of the other figures.

[a] Test not evaluated individually, but is included in a combination or composite score:
[a1]pilot composite; [a2]working memory span composite; [a3]speed of processing; [a4]cognition battery; [a5]AFOQT composite
[b] Domain is listed under the test referenced in the description.

Appendix I. Article Appraisal, Selected Criteria

Table I-1: Rating of appraisal criteria for included articles (rows shaded = met all criteria)

First Author, Year	Sample Size >30	Predictive or comparative study design	Confounding addressed in any way	Categorized Subjects	Total criteria met
Barron, 2016	+	+	-	-	2
Basner, 2015	-	+	-	-	1
Bell, 1997	+	+	-	-	2
Caretta, 1996	+	+	+	-	3
Causse, 2010	-	+	+	-	2
Causse, 2011a	-	+	+	+	3
Causse, 2011b	+	+	+	+	4
DeVoll, 2013	+	-	-	+	2
Doane, 2003	+	-	-	-	1
Endsley, 1990	-	+	-	-	1
Endsley 1994	-	-	-	-	0
Griffin, 1987	-	+	+	-	2
Hyland, 1994	+	-	-	-	1
Kay, 1995 CogScreen manual Phase C Clinical	+	+	?	+	3
Kay, 1991 + CogScreen manual Phase B study	+	+	+	+	4
Kay, 1993 + CogScreen manual US/Russian norms	+	+	+	-	3
Kennedy, 2010	+	+	+	-	3
Kennedy, 2013	+	+	+	-	3
Kennedy, 2015	+	+	+	-	3
Le Roux, 1988	+	+	-	+	3
McGuire, 2014	+	+	-	-	2
McGuire, 2016	+	+	+	-	3
Moore, 1996	-	-	-	-	0
Morrow, 2003	+	+	+	-	3
O'Donnell, 1992	+	+	+	+	4
Shull, 1990	+	-	-	-	1
Stokes, 1991	+	+	+	+	4
Stokes, 1992	-	+	+	+	3
Sulistyawati, 2011	-	+	+	-	2
Taylor, 2000	+	+	+	+	4
Taylor, 2005	+	+	+	-	3
Taylor, 2007	+	+	+	-	3
Tolton, 2014	+	+	+	+	4
Van Benthem, 2016	+	+	+	-	3
Venturino, 1990	-	-	-	-	0
Waag, 1994	+	-	-	-	1
Yakimovich, 1994 + CogScreen manual	+	+	+	+	4
Yesavage, 2011	+	+	+	-	3
Zhang, 1997	+	+	-	+	3

+ criteria met; - criteria not met; ? unclear

Appendix J. Transcripts of Conference Presentation

(Additional Clarification was obtained from presenter and added to transcript)

1. Development of Aviator Norms for CogScreen for CogScreen presented by Gary Kay. Aerospace Medical Association's 64th Scientific Meeting held in Toronto, Canada 1993

2. Flight Performance and CogScreen test battery in Russain Pilots presented by Gary Kay. Aerospace Medical Association's 65th Scientific Meeting held in San Antonio, Texas 1994

---

1. Development of aviator norms for Cogscreen presented by Gary Kay, Aerospace Medical Association's 64th Scientific Meeting held in Toronto, Canada 1993

Kay G, Strongin G, Hordinsky J, et al. Development of aviator norms for Cogscreen. Aerospace Medical Association 64th Annual Scientific Meeting; 1993 Toronto, Canada.

Power of Development and Aviator News for CogScreen

Being presented by Gary Kay. His co-authors are Gregory Strongin, Jerry Hordinsky and Barton Pakull.

> Well, I should start off by saying: This is a collaborative project. I am from Georgetown University, Dr. Gregory Strongin is from the Russian State Research Institute of Civil Aviation and apologizes for his inability to be here with us today. He had planned to but, wasn't able to join us. Dr. Jerry Hordinsky is from the FAA Civil Aviation Medical Institute and Dr. Bart Pakull is from the FAA's Office of Aviation of Medicine. And, I also want to acknowledge the very important contributions from Nadia Yakimovich who is responsible for most of the testing in Moscow and Vitali Govoreshenko who is also involved in the testing and technical aspects of the testing done in Moscow. Alexander Chervinsky was my Post Doctoral Fellow, and was responsible primarily for the translation of CogScreen into Russian, and Sarah Morris, who assisted me at Georgetown. Alan Stokes who helped pave the way by making the *first* trip over to determine the feasibility of doing this research program in Moscow.
>
> First of all, let me begin with the history of CogScreen. CogScreen was developed in response to the FAA's requirement for a cognitive screening test that could detect subtle levels of brain dysfunction that if left unnoticed could potentially interfere with skilled aviation performance. We developed CogScreen in response to this FAA requirement. The FAA was seeking a 45-60 minute self-administered cognitive screening test.

CogScreen, is administered with a light pen. [It is now performed with a stylus on a touchscreen.] All of the examinees answers and responses are made with a light pen, directly onto the screen, except for the tracking test which uses 2 keys on the keyboard.

This is the test battery menu. If you press for example, the top choice, the entire battery is administered. The tests, as I said before, were completely translated into Russian for this project.

This is a Backward Digit Span subtest. Next is the Visual Sequence Comparison subtest, a sequence comparison test where there are numbers and letters like F4DIL and F0DIL – the examinee looks at these and decides: Same or Different.

This is the Matching to Sample subtest You would look at that pattern for a few moments. It disappears and then you would have two patterns. With the light pen you'd press the pattern that matched the one previously shown in the center.

This is the Manikin subtest in which you determine which hand is holding the flag and press the corresponding left or right box.

This is a Symbol Digit Coding subtest.

This is the immediate recall trial for the Symbol Digit Coding subtest, a paired associate learning task.

There's a subtest called the Pathfinder test. In the corners of the screen boxes are shown containing either a number or a letter. In the first version, the Numeric Pathfinder subtest, there are numbers in the boxes. The examinee presses the box with the next number that would occur in sequence. So, if you just pressed 2, you'd look for the box containing 3 and press that box with the light pen. If you were doing the alphabetic sequence, you would look for the next letter in sequence. If you were doing the combined sequence you would following an alternating sequence of numbers and letters. The Pathfinder subtest is an analog of the Trail Making Test.

There is a Divided Attention subtest where you perform a visual monitoring task simultaneously with the visual sequence comparison test.

The Shifting Attention Test involves various different cognitive tasks. For the Border Color condition, you would match to the color of the border. This is one of the 3 response rules. For the other parts of the Shifting Attention Test you have to apply or determine by deductive reasoning which is the active rule.

The last task in CogScreen is the Dual Task subtest which involves both a tracking test and a previous number exercise. For the previous number task you're shown a number, which disappears and is replaced by a new number. Your task is to tap the previous number shown. If the number shown in the box was a 2 and was replaced by a 3 you would press the 2 with the light pen. If a 1 came up after the 3, you would press the 3. So, you're always responding with the number previously shown.

The FAA supported the large scale normative data collection in the United States. We tested over 640 commercial airline pilots. The sensitivity and specificity of

CogScreen was determined to be quite good for determination or detection of brain dysfunction. We've presented that information elsewhere. The FAA decided to share the CogScreen technology with the Russian State Research Institute of Civil Aviation through an existing program of collaborative research. The aim of the collaborative effort was to obtain additional normative data on CogScreen, on healthy airline pilots and to compare the normative data between the two countries. This was an opportunity to look at cross cultural factors involved in CogScreen performance. Another goal was to determine the suitability of CogScreen for future use in the medical evaluation of Russian pilots. The Russians indicated a need for a fitness-for-duty type instrument. A testing protocol, which duplicated the one used for normative data collection in the United States, was implemented in the laboratory of the State Research Institute at the Vnukovo Airport outside of Moscow. The protocol included completion of an informed consent form, which was translated into Russian. (That was one of the harder adjustments for the Russians; this use of an informed consent form). And a demographic and medical data form. There was initially an objection to asking questions about alcohol use because if they had alcohol problems, they wouldn't be a pilot. Our Russian colleagues later decided that they would include the alcohol questions. CogScreen was administered at Vnukovo Airport, Moscow, in a quiet and private office. The version of CogScreen which was administered was fully translated into Russian and is identical to the U.S. version in all other respects.

Demographic and performance data were encoded and analyzed using SPSS-PC. An analysis of variance was performed, with and without co-variants to control for effects of age, computer experience, hours of sleep and alcohol use. These factors previously had been determined to affect performance on CogScreen. Demographic factors were compared and the correlations between these demographic factors in CogScreen were calculated. We also performed discriminant analyses and multiple regression analyses.

Comparison of the pilot groups on the basis of demographic variables shows that the two groups were well-matched with respect to total flight hours, age and sex. The Russian pilots appeared to be slightly more sleep-deprived with an average of 6.8 hours of sleep prior to CogScreen, compared to 7.1 hours for their American counterparts. Subjective ratings of sleepiness (using the Stanford Sleepiness Scale) were similar for the two groups. The frequency with which pilots admitted to becoming inebriated during a 12-month period prior to the testing did not differ between the two groups. However, the quantity of alcohol consumed was greater for the Russian groups. At those times in which they said they'd become inebriated, Russian pilots admitted to consuming more alcohol. Russian pilots who reported heavy alcohol consumption reported to be drinking in excess of 250 grams of alcohol – approximately a half bottle of vodka at a sitting. By comparison, American pilots who said they drank heavy reported that they tended to drink about 125 grams or a quarter bottle equivalent of vodka. That's translating alcohol use into vodka units, so to speak.

The most dramatic difference, as you can see from looking at the table, was in computer literacy – where we had pilots rate themselves from Novice (1) to Experienced Computer User (4). As you can see, at the Novice (or No use) level, we have 25 percent of the U.S. group compared to 66 percent of the Russian group. At the Limited use level, we found 27 percent and at Level 3, which reflects moderate computer use, we found 32 percent of the American group. By comparison only 6 percent of the Russian group reported Moderate use and only 1 percent identified themselves as particularly skilled users of computers.

Comparing the 203 Russian pilots with the 586 U.S. pilots, there was a great deal of similarity in CogScreen results but also some differences. We have 58 different dependent measures that we compared for the two groups. There were 20 different speed measures, 20 accuracy measures, and in addition Thruput measures which truly are not an independent group of measures. These are measures derived from speed and accuracy. Basically, they reflect the number of correct responses per minute. And so, they include both speed and accuracy. We've focused on non-redundant variables in performing this analysis.

What I've done is, rather than show the number of significant F test – because when you have 587 subjects in one group and 203 in the other, I can assure you, you get a lot of $p<.05$ differences. Those considered meaningful, were those where the group difference accounts for more than 9 percent of the variance. I've shown here the percentage of variance accounted for by the comparison. This table shows differences in CogScreen performance based on pilot nationality. The largest difference was found on the letter sequencing task (i.e., Pathfinder Letter). Now, even though this subtest was presented using Cyrillic letters the Russian pilots, still had significantly weaker performance with both accuracy and speed on this letter sequencing task.

Another area where you see differences is on the Auditory Sequencing subtest, a measure of sound pattern comparison. The examinee listens to a sequence of tones followed by a second sequence of tones and decides whether the two sound patterns were the same or different. The Russians had considerably more difficulty in performing this task. With regards to accuracy, the response accuracy for Russian pilots was 84 percent compared to 93 percent for U.S. pilots. Nationality accounted for 16 percent of the variance in task performance. There were some other differences, most of which relate to tracking and speed on tasks involving letter processing. The Russian pilots significantly outperformed the US pilots on the Math subtest, though nationality accounted for less than 9 percent of the variance. For the Math subtest the examinee reads a math word problem and then selects the correct answer from among three choices. As you can see, the greatest number of differences *were* on response speed measures.

Further analyses showed significant age effects on most of the speed variables. Age effects were also found on some response accuracy measures, particularly on the Auditory Sequence Comparison and Backward Digit Span subtests. Computer use and experience played an important role in performance on the Pathfinder Letter subtest and Math subtest. The effect of computer use, and experience is highly associated with Nationality. For the Math subtest Thruput declined as a

function of higher computer use. In contrast, for the letter sequencing task, computer use was positively associated with better test performance. For example, Russians with the least amount of computer use were those who performed best on the Math test. In fact, for this subtest, the effect due to nationality was completely eliminated when the analysis of covariance controlled for computer use. Perhaps, mental math skills weaken with the use of these laptops.

Years of education was correlated with performance on the Pathfinder subtest (letter sequencing), Math, and the Previous Number task. On the other hand, a history of heavy alcohol use, that is excessive use, 250 grams or greater use, was associated with poorer accuracy on the Auditory Sequence Comparison subtest. An interesting finding, considering that our results for 50 recovering alcoholics in the US, also showed poor performance on this Auditory Sequence Comparison subtest.

Another subtest impacted by heavy alcohol use was the Shifting Attention Test Arrow Direction subtest. Frequency of inebriation during the course of 12 months was associated with performance on Pathfinder Combined and the Visual Sequence Comparison subtests.

The analysis shows that we could account for approximately 50 percent of the variance attributable to nationality with those seven CogScreen variables. This was shown with both discriminant analysis and multiple correlation analysis.

With the discriminant analysis we correctly identify nationality for 89 percent of the pilots. The classification analysis shows that 455 of the U.S. pilots and 161 of the 203 Russian pilots were identified correctly (as Russian pilot) based on just those seven variables.

In conclusion, both Russian and U.S. colleagues, working with CogScreen are of the impression that CogScreen would be a useful tool in the assessment of pilots requiring evaluation following changes in their neurologic, psychiatric and / or performance status. We believe that there are explanations for most of the differences that we've seen here in CogScreen, comparing US and Russian airline pilots. Of the 58 variables generated by CogScreen, there were 7 that differentiated the two groups.

The poorer performance on letter sequencing by Russian pilots appears to be due to the fact that Russians have less familiarity with the last half of their alphabet. The Combined Pathfinder subtest includes letters and numbers, up through the first 12 letters of the alphabet. The Russian pilots performed equally well as the American pilots on this task. In contrast, Russian pilot performance was significantly worse than US pilots on the Pathfinder Letter subtest which requires sequencing of 25 letters. They appeared to have difficulty sequencing the last half of their alphabet. The head physician of the Russian Civil Aviation Administration, Dr. Khavatov, commented that the Russians should commission the development of an alphabet song similar to the popular American alphabet song, i.e., the "A, B, C song".

We offer two possible explanations for the large difference in accuracy on the Auditory Sequence Comparison subtest – One possibility is that there may be a higher level of ambient noise in the Russian commercial aviation cockpit. We do not have the noise level values for the different cockpits. The other explanation is that the difference may be related to alcohol use.

The difference in Math speed, as I indicated, appears to be accounted for by computer experience. Another factor may be how math is taught in the Russian education.

Differences in training and selection of pilots may be responsible for slower reaction times and poorer tracking performance. I was surprised to learn that Russian airline pilots are solely civil aviation pilots without any military aviation background. In Russia, if you want to be a pilot you go into either civil aviation or military aviation. You don't cross over. You don't retire from military aviation and go into civil aviation. So, the Russian pilot sample doesn't have the high number of jet fighter pilots (and other military aviators) who are in the U.S. pilot sample.

I'll end by saying that the Russians favorable regard for CogScreen is indicated by their use of CogScreen in a number of biomedical research studies, including studies on anti-hypertensive medications, head injury studies, and studies investigating patients with thyroid hormone problems. The Russians are currently developing protocols to look at the relationship between job performance – actual cockpit performance in commercial aviators in Russia and relating that to their CogScreen performance and to other job readiness measures. Thank you very much.

---

2. Flight performance and CogScreen test battery in Russian pilots presented by Gary Kay Aerospace Medical Association's 65th Scientific Meeting held in San Antonio, Texas 1994

Yakimovich NV, Strongin GL, Govorushenko V, et al. Flight performance and CogScreen test battery in Russian pilots. Aerospace Medical Association 65th Scientific Meeting San Antonio Texas 1994.

Personality and Factors Predicting Performance of Aerospace Personnel

Co-chair Dr. Carol Manning for this session entitled "Personality, Cognitive, and Other Factors Predicting Performance of Aerospace Personnel," and just as a reminder, the presentations will be limited to about ten minutes with approximately five minutes for questions and any attendees that have questions, you can use the mike that's in the center of the main aisle, center aisle, and with that, our first paper is – do you want to introduce that?

Okay. I get to try to pronounce these names. The first paper is entitled "Flight Performance and CogScreen Test Battery in Russian Pilots." The authors are Nadia Yakimovich, Gregory Strongin, Vitaly Govorushenko, all from the Russian State Institute of Civil Aviation, Dave Schroeder from the Civil Aviation Medical Institute, and Gary Kay who will be doing the

presentation, who is from Georgetown University School of Medicine. Let him pronounce the names over again so you will be able to hear how they really sound.

Gary Kay: I'll actually let you look at the names and try to pronounce them to yourselves. This research investigated CogScreen as a predictor of flight performance in Russian pilots. The first three authors are from the Russian State Research Institute of Civil Aviation, Nadia Yakimovich, Gregory Strongin, and Vitali Govorushenko. Dr. Schroeder is from CAMI and I'm from Georgetown.

CogScreen is a fully computerized test battery that was developed for use in medical certification of airmen. The test was designed to meet the FAA's need for a standardized low-cost instrument that could detect subtle levels of brain dysfunction that could potentially interfere with pilot performance. Validation studies demonstrated that CogScreen has sensitivity and specificity in evaluation of mild brain dysfunction. We've collected normative data with CogScreen on over 550 U.S. Pilots and through a collaborative U.S.-Russian joint research program we have normative data on over 200 Russian pilots tested with a Cyrillic version of CogScreen that was made available to the Russian State Research Institute. Last year at this meeting, we presented a comparison of the Russian and American normative data. This year's presentation is focused on determining the relationship between CogScreen and actual flight performance of pilots.

The Russia State Research Institute of Civil Aviation through an arrangement with AeroFlot obtained access to flight performance logs. These logs list for each captain the flight parameter violations registered by the aircraft's flight data recorder, the "aircraft's Black Box." The flight data recorder was analyzed by a special computer program that generates a report for each flight indicating whether any of 64 different flight parameter violations has occurred. This analysis program was developed for the purpose of evaluating and training Russian airline pilots. Examples of the types of flight parameter violations registered by the flight data recorder program are shown here on this slide. We tested a total of 75 captains with a mean age of 46.6 years. Performance data was available for a three-year period for each of these captains. The pilots in this study flew either the IL-86 which is shown on the bottom of the slide. This is a four-engine under-the-wing aircraft similar to a DC-8. It's used in transcontinental aircraft operations. The airplane shown above, the TU-154, is similar to our DC-9 or B-727 aircraft and is used heavily in domestic flight operations.

The seriousness of each of the 64 different flight parameter violations was determined by ratings that were developed by five very senior captains with Aeroflot. Each captain was asked to rate the overall seriousness of each violation on a three-point scale with one being least serious, three being most serious. The inter-rater reliability was 0.81. These ratings were then used to develop an Index of Flight Performance for each pilot. The sum of the violations weighted by severity were divided by the pilot's total number of flight hours over the three-year period. This sum was then multiplied by 100 to obtain the final Index of Flight Performance score.

Pilots were administered the Cyrillic (i.e., Russian language) version of the CogScreen test under standard conditions. Testing required approximately one hour. Subjects made their responses on this test using a light pen. To remind people about CogScreen, I'll just cover a couple of the sub-tests. There's a Backward Digit Span subtest where the person is presented with digits, like here it would be an 8, and another digit would come up, a 5, and then the examinee's task is to remember those digits in reverse order. The examinee responds by tapping the light pen the digits shown at the bottom of the screen (in reverse order). The Backward Digit Span subtest is a

measure of visual attention span (and numeric working memory). Next, I'll show the Matching to Sample subtest. The examinee is presented a checkerboard pattern. The pattern disappears and is replaced by two checkerboards. The examinee presses the light pen to the checkerboard that matches the one previously shown. Another subtest, one of our two divided attention measures, presents a visual monitoring task. The examinee observes a cursor traveling up and down in a circle. When the cursor enters a blue area, the examinee presses the Center key with the light pen. That returns the cursor to the center of the circle. Simultaneously the examinee compares sequences of numbers and letters, like JM6AP, JM6AP. If they're the same, they press Same. If they're different, they press Different. Another task, which is more conceptually demanding, involves application of simple response rules. For example, under one condition the examinee is instructed to select a box that matches in terms of the border color. One task involves reading and applying the response rule. Another task involves using deductive reasoning to discover the active rule.

My Russian co-authors used the Index of Flight Performance to classify pilots into one of three groups: optimal performance, adequate performance, and sub-optimal performance. Membership in the optimal performance group included all pilots who did not have any flight performance violations or deviations during the course of the three-year recording period. The index score for the adequate performance group differed according to aircraft. For the IL-86, the larger airplane, the adequate range was an index score from 1 to 8.06. For the 154 aircraft, the adequate range was 1 to 3.87. Scores exceeding the adequate range fell into what was classified as sub-optimal performance.

Correlations were performed to assess the relationship between CogScreen and the Index of Flight Performance with and without controlling for the effects of age. Multiple regression analysis of CogScreen variables and the Index of Flight Performance were also performed.

Results show that the mean age of the TU-154 pilots was 45 years. The mean age for IL-86 pilots was 49. This difference was statistically significant. The age distribution for the pilots flying the two aircraft also differed significantly. As you can see here, for the TU-154 pilots the distribution shows similar numbers of pilots across age groups. This is not the case, for the IL-86 where a large number of pilots fall in the 46 to 50 and 41 to 45 year age group. For this IL-86, 44 percent of the pilots were in the 51 to 55 year age group and another 24 percent in the 46 to 50 age range.

There was a significant correlation between performance on the Index of Flight Performance and age for the TU-154 pilots, but not for the IL-86 pilots. These figures show the distribution of Index of Flight Performance scores for the two aircraft. The figures show a more skewed distribution for the TU-154. Fifty-four percent of the TU-154 pilots did not have any violations over the 3 years compared to eight percent of the IL-86 pilots over the same time period. At the other end of the scale, 32 percent of the IL-86 pilots had performance scores of nine or greater compared to only six percent of the TU-154 pilots. The average Index of Flight Performance scores for the two aircraft were significantly different, a mean of 2.3 for the TU-154 and 6.4 for the IL-86. These outcomes clearly suggest that there are differences in performance of pilots flying these two aircraft. However, we don't have sufficient information to determine the extent to which these differences can be attributed to how pilots are assigned to the two aircraft, characteristics of the aircraft, or possible differences in how the flight data recorder on the two aircraft sample the various flight parameters.

The distribution of pilots falling in the Optimal, Adequate, and Sub-optimal performance levels, also differed for the two aircraft. More TU-154 pilots were in the Optimal group. For the IL-86 there were more pilots in the Adequate group than for the TU-154. We found approximately the same percentage of pilots for both aircraft in the Sub-optimal group.

Analysis of the correlations between CogScreen and the Index of Flight Performance show a number of significant correlations for the TU-154 and the IL-86. Nine CogScreen variables exhibited significant correlations with flight performance on the TU-154. Seven CogScreen variables remained significant when the effect of age was partialled out. CogScreen variables that were significantly correlated with the Index of Flight Performance included the Divided Attention Test speed under the dual task condition. The Dual Task Test Previous Number test under the dual task condition was also correlated. The Matching to Sample speed variable, a measure of visual working memory, and the Shifting Attention Test Discovery condition, a measure of deductive reasoning, were correlated with the Index of Flight Performance.

Three of the seven variables showing correlations between the TU-154 Captains and the Index of Flight Performance, were also significantly correlated for the IL-86; including the Divided Attention Test speed variable, the Dual Task Test Previous Number (dual task) Thruput and speed. In addition, for the IL-86, we found a significant correlation of r=0.46 for the Backward Digit Span subtest, which is a measure of numeric working memory. Correlations between the Index of Flight Performance and CogScreen variables were virtually identical with and without controlling for the effect of age.

Results from the multiple regression analysis using the CogScreen variables to predict flight performance for the two aircraft are shown here. Multiple regression analyses revealed that four CogScreen variables resulted in a multiple R of 0.61 and an R^2 of 0.30 for the TU-154 aircraft. Divided attention, conceptual reasoning, and tracking performance, especially under multi-tasking conditions, were predictive of the performance of pilots flying the TU-154. For the IL-86, we have a different group of predictive variables and the prediction was somewhat better than for the TU-154. Three variables resulted in a multiple R of 0.73 and an adjusted R^2 of 0.45. For IL-86 pilots, performance was predicted by their scores on CogScreen measures of visual scanning and sequencing, concentration, and divided attention.

Performance on CogScreen variables for the three performance groups (Optimal, Adequate, and Sub-Optimal) by aircraft, with age as a co-variant, are presented in the next slide. The analysis shows that the performance group assignments into the Optimal, Adequate, and Sub-optimal groups were significantly different for the two aircraft independent of age.

Since this is a retrospective study, some caution should be exercised in the interpretation of the results. This is especially true given the differences in age and performance of Captains assigned to the two aircraft included in the study. Additional information is needed concerning the flight characteristics of the two aircraft and policies associated with how pilots are assigned to these two aircraft. An additional caution is called for by our use of an Index of Flight Performance score where performance violations are weighted such that the same index score can be obtained by having a number of minor violations or by having a smaller number of more serious (highly weighted) violations.

Despite differences in the performance of Captains assigned to these two aircraft, there are indications that aspects of CogScreen performance are predictive of pilot performance (i.e., frequency and severity of flight violations). While generally modest in magnitude, several of the

correlations between individual CogScreen variables and performance of aviators in this sample were significant. The most highly correlated variable for predicting performance for the IL-86 Captains was a divided attention task, which accounted for approximately 30 percent of the variance in the flight performance index score. The next most highly correlated variables included a measure of numeric working memory (i.e., Backward Digit Span), and additional measures of multitasking. Correlations between CogScreen variables and performance of the TU-154 Captains were generally lower. However, there was still a large number of significant correlations. Although, the correlation between individual CogScreen variables and the Index of Flight Performance was modest, the multiple regression findings were far more impressive. CogScreen variables accounted for 45 percent of the variance in the Index of Flight Performance for the IL-86 Captains. Four of the CogScreen variables accounted for 30 percent of the variance in the flight performance index for the TU-154 Captains. It's particularly interesting to note the difference in the variables predictive of performance for these two aircraft.

Finally, while these results are promising, additional research is needed to more clearly demonstrate the relationship between various CogScreen variables and flight performance. The present study was retrospective. Next year, we have to report to you on a prospective study relating CogScreen performance with six indices of approach and landing quality taken from a series of landings under a defined set of flying conditions. Additional information will be included concerning pilot perceptions of their performance and assessment of other factors that potentially influence flight performance. We also want to point out that this study demonstrates the mutually beneficial nature of our Russian-American collaborative research program. Thank you.

Appendix K. Evidence Tables

Detailed Evidence Tables are available as supplemental materials from the authors