AFRL-RH-WP-TR-2024-0029

AIR FORCE OFFICER QUALIFYING TEST (AFOQT) FORM U: REVIEWING PSYCHOMETRIC PROPERTIES OF CANDIDATE COGNITIVE SUBTESTS

Julia L. Walsh, Rusty Wilson, Kyle J. Mann,
Shane Sizemore, and Montana R. Drake
DCS Corp

John Trent
Air Force Personnel Center
Strategic Research and Assessment Branch
JBSA Randolph, TX

Thomas R. Carretta
Air Force Research Laboratory
Performance Optimization Branch
Wright-Patterson AFB, OH

June 2024
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2024-0029 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION.

CARRETTA.THOMAS.R.1228984929 Digitally signed by CARRETTA.THOMAS.R.1228984929 Date: 2024.06.06 13:16:01 -04'00'

THOMAS R. CARRETTA, PhD
Work Unit Manager
Performance Optimization Branch
Air and Space Biosciences Division

WILLIAMS.LOGAN.ANDREW.1273597634 Digitally signed by WILLIAMS.LOGAN.ANDREW.1273597634 Date: 2024.06.07 11:17:18 -04'00'

LOGAN R. WILLIAMS, DR-III, PhD
Human Performance Area Lead
Operational Product Section
Product Development Branch
Air and Space Biosciences Division

This report is published in the interest of scientific and technical information. And its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (*DD-MM-YY*)	2. REPORT TYPE	3. DATES COVERED (*From - To*)
01-06-24	Interim	January 2022 – May 2022

4. TITLE AND SUBTITLE
Air Force Officer Qualifying Test (AFOQT) Form U: Reviewing Psychometric Properties of Candidate Cognitive Subtests

5a. CONTRACT NUMBER
FA8650-21-F-4104

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Julia L. Walsh[a], Rusty Wilson[a], Kyle J. Mann[a], Shane Sizemore[a], Montana R. Drake[a], John Trent[b], and Thomas R. Carretta[c]

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H12Q

7. PERFORMING ORGANIZATION N/AME(S) AND ADDRESS(ES)
DCS Corp[a]
4027 Colonel Glenn Highway, Suite 210
Dayton, OH 45431-1672

8. PERFORMING ORGANIZATION REPORT NUMBER

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

Air Force Materiel Command
Air Force Research Laboratory[c]
711th Human Performance Wing
Airman Systems Directorate
Air and Space Biosciences Division
Performance Optimization Branch
Wright-Patterson AFB, OH 45433

Air Force Personnel Center[b]
Strategic Research and Assessment Branch
550 C Street West, Ste. 10
JBSA Randolph AFB, TX 78150

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHBC

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-TR-2024-0029

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

13. SUPPLEMENTARY NOTES
Report contains color. AFRL-2024-3041, cleared 6 June 2024

14. ABSTRACT
The Air Force Personnel Center Strategic Research and Assessment Branch (AFPC/DSYX) initiated a contract to evaluate cognitive subtests not previously included in the Air Force Officer Qualifying Test (AFOQT) to determine which ones should be considered for inclusion in the next generation AFOQT Form U. The current effort reviewed 16 candidate Department of Defense (DoD) batteries and subtests. Archival data were available for some subtests. Of those, two were recommended for Form U on an experimental basis – Weight Perception (WPt) and Non-Verbal Reasoning (NVR). The subtests for which the data were not available were reviewed based on the extant literature. Of those, two were recommended for inclusion in Form U, also on an experimental basis – Assembling Objects (AO) and Mental Counters (MCt). The rest of the tests were either already included in the AFOQT or had psychometric deficiencies, such as lack of incremental validity or high mean score subgroup differences. We also examined criterion-related validity for Physical Science (PS) across four samples. PS has been included in the AFOQT Form T since 2015 on an experimental basis. Given the evaluations, it is unlikely that PS would add incremental prediction above and beyond the other AFOQT subtests for officer commissioning or aircrew training qualification. Additional analyses are needed to determine whether PS has utility for predicting performance in non-aircrew training courses. Finally, we reviewed available literature to determine what other constructs should be considered for inclusion in AFOQT Form U. Based on our review, we recommended that the AFOQT content be expanded to include measures of inductive reasoning (fluid intelligence), learning/learning agility, short term (working) memory, long term (meaningful) memory, and emotional intelligence.

15. KEY WORDS
Air Force Officer Qualifying Test, Assembling Objects, Non-Verbal Reasoning, Mental Counters

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. N/AME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Thomas R. Carretta
Unclassified	Unclassified	Unclassified	SAR	58	19b. TELEPHONE NUMBER (*Include Area Code*) N/A

Table of Contents

LIST OF TABLES

1.0 EXECUTIVE SUMMARY

The Air Force Personnel Center Strategic Research and Assessment Branch (AFPC/DSYX) initiated a contract to evaluate cognitive subtests, not previously included in the Air Force Officer Qualifying Test (AFOQT) to determine which should be considered for the inclusion in the next generation AFOQT Form U. This project is part of a continuous effort by the United States Air Force (USAF) to predict important personnel outcomes while increasing qualification rates for historically underrepresented gender, racial, and ethnic minority subgroups (a phenomenon known as the diversity-validity dilemma; Ployhart & Holtz, 2008).

This project stemmed from a previous effort wherein an empirical, theoretical, and competency evaluation of the 17 AFOQT cognitive subtests that appeared in previous forms was conducted (see Walsh, Wilson, et al., in press). The results of that effort revealed that the AFOQT Form U should reimagine its suite of cognitive subtests to include measures of fluid intelligence and/or specific abilities. The current effort reviewed 16 candidate Department of Defense (DoD) batteries and subtests thar purported to measure those constructs.

The primary goal of this review was to identify subtests that would complement the AFOQT Form T cognitive subtests (Verbal Analogies [VA], Arithmetic Reasoning [AR], Math Knowledge [MK], Reading Comprehension [RC], Physical Science [PS], Table Reading [TR], Instrument Comprehension [IC], Block Counting [BC], and Aviation Information [AI]). AFOQT Form T also includes two non-cognitive subtests (Situational Judgment Test [SJT] and Self-Description Inventory - Officers [SDI-O)]. Archival data were available for some subtests. Of those, two were recommended for Form U on an experimental basis – Weight Perception (WPt) and Non-Verbal Reasoning (NVR). The subtests for which the data were not available were reviewed based on the extant literature. Of those, two were recommended for inclusion in Form U, also on an experimental basis – Assembling Objects (AO) and Mental Counters (MCt). The rest of the tests measured the constructs that were either already measured by the AFOQT Form T subtests or had psychometric deficiencies, such as lack of incremental validity or high mean score subgroup differences.

We also examined criterion-related validity of PS, an experimental subtest included in the AFOQT Form T. Four samples were used – Officer Training School (OTS) sample, Pilot sample, Air Battle Manager (ABM) sample, and Combat Systems Officer (CSO) sample. In the OTS sample, PS showed modest correlations with academic-based criteria but did not provide practical incremental variance above and beyond the existing AFOQT academic composite. In the Pilot sample, PS did not show significant correlations with the training criteria. In the CSO sample, PS showed moderate correlations with training criteria and provided little incremental variance (1 percent (%) or less) above and beyond the existing CSO composite. In the ABM sample, PS showed moderate correlations with training criteria but did not provide incremental variance above and beyond the current ABM composite. Given our evaluation, we do not expect that PS would add practical incremental prediction above and beyond the other AFOQT subtests for officer commissioning or aircrew training qualification; however, more research is needed to determine if PS can increment prediction for non-rated career fields (e.g., cyber, intelligence).

The secondary goal of the current effort was to review available literature to see what other

constructs should be considered for AFOQT Form U. Based on our evaluation, we recommended that AFPC/DSYX consider the inclusion of subtests that measure inductive reasoning (fluid intelligence), learning/learning agility, short term (working) memory, long term (meaningful) memory, and emotional intelligence. The literature review also revealed perceptual/processing speed as one of the key constructs. However, the AFOQT already includes a direct measure of perceptual/processing speed, called TR. The literature also revealed some soft skills that may need to be assessed. We recommended that they be measured by a different instrument than the AFOQT, for example a biographical data (biodata) method.

Reviewing the literature and the experimental subtests that we recommended for inclusion in From U, some of the existing DoD subtests may capture the missing constructs. Specifically, MCt, AO, NVR, and WPt may capture fluid intelligence and working memory.

In sum, this effort and others (e.g., Kantrowitz et al., 2023; Walsh, Brady, et al., 2022; Walsh, Wilson, et al., in press; Walsh, Woolley, et al., 2022) laid the foundation for AFOQT Form U development. The current Form U blueprint includes nine Form T cognitive subtests (VA, AR, MK, RC, PS, TR, IC, BC, and AI), two non-cognitive subtests (SJT and SDI-O), and four experimental subtests (AO, NVR, WPt, and MCt). The expectation is that the experimental subtests will capture some of the missing constructs (thereby expanding the competencies measured in commissioning officers and officers pursuing rated career fields) and also increase aperture for historically underrepresented gender, racial, and ethnic subgroups. The next step in the project entails generating a large pool of items for each subtest and collecting data for initial psychometric evaluation.

2.0 INTRODUCTION

The purpose of this project was to provide AFPC/DSYX researchers and policy makers (Air Force Accession Policy, AF/A1PT;) evidence-based recommendations regarding the content of the next-generation AFOQT Form U. This project is part of a continuous effort by the Air Force to improve prediction of important personnel outcomes while increasing qualification rates for historically underrepresented gender, racial, and ethnic subgroups (a phenomenon known as the diversity-validity dilemma; Ployhart & Holtz, 2008). The activities described in this technical report stemmed from a previous effort wherein an empirical, theoretical, and competency evaluation of 17 AFOQT cognitive subtests from previous forms was conducted (see Walsh, Wilson, et al., in press). The next section briefly describes that effort.

2.1 Reviewing Cognitive Subtests from Previous Forms

The first step in developing the AFOQT Form U involved an evaluation of the 17 cognitive subtests that appeared in previous AFOQT forms (O, P, Q, R/S, and T). These subtests were either currently operational (Form T) or previously operational (now de-commissioned). Considering the results of the evaluation, it was recommended that the next generation AFOQT Form U (1) excludes the de-commissioned subtests due to their poor psychometric properties; (2) removes one Form T subtests called Word Knowledge (WK) due to its poor psychometric properties and weak linkage to the USAF competencies; and (3) updates the content of the rest of the operational subtests. See Table 1 for a list of cognitive subtests that appeared in one or more forms between 1980 and 2015 (Forms O through T).

Table 1. Summary of the 17 AFOQT Cognitive Subtests

Subtest	Recommendation*	Cattell-Horn-Carroll Theory of Intelligence*	Factor Analysis**	Subgroups that Showed Moderate to Large Mean Score Subgroup Differences**	Test Length and Administration Time***
VA	Include	Comprehension Knowledge, Reading	General mental ability (*g*)	Black/White	25 items and 8 min
AR	Include	Quantitative Knowledge	*g*	Black/White Female/Male	25 items and 29 min
WK	Not include	Comprehension Knowledge	*g* Verbal Ability	Black/White Asian/White	25 items and 5 min
MK	Include	Quantitative Knowledge	*g*	Black/White	25 items and 22 min
RC	Include	Reading	*g*	Black/White Female/Male Asian/White	25 items and 38 min
PS	Include (conduct additional analyses to examine utility for non-rated career fields)	Domain-Specific Knowledge	*g*	Black/White Female/Male	20 items and 10 min
TR	Include	Fluid Reasoning, Processing Speed	*s* Perceptual Speed	Black/White	40 items and 7 min
IC	Include	Domain-Specific Knowledge, Visual-Spatial Processing	*g* and Aviation Knowledge	Black/White Female/Male Asian/White	25 items and 5 min
BC	Include	Fluid Reasoning, Visual-Spatial Processing	*s* Spatial Ability	Black/White Female/Male	30 items and 4.5 min
AI	Include	Domain-Specific Knowledge	*g*	Black/White Female/Male Asian/White	20 items and 8 min

EM	Not include	Fluid Reasoning, Visual-Spatial Processing	*Not Available*	*Not Available*	20 items and 10 min
GS	Not include	Domain-Specific Knowledge	*Not Available*	*Not Available*	20 items and 10 min
HF	Not include	Fluid Reasoning, Visual-Spatial Processing	*Not Available*	*Not Available*	15 items in 8 min
MC	Not include	Domain-Specific Knowledge, Fluid Reasoning	*Not Available*	*Not Available*	20 items in 22 min
DI	Not include	Fluid Reasoning	*Not Available*	*Not Available*	25 items and 24 min
RB	Not include	Fluid Reasoning, Visual-Spatial Processing	*Not Available*	*Not Available*	15 items in 13 min
SR	Not include	Domain-Specific Knowledge, Fluid Reasoning	*Not Available*	*Not Available*	40 items in 15 min

Note. Underlined are the subtests that appear on AFOQT Form T. VA = Verbal Analogies; AR = Arithmetic Reasoning; WK = Word Knowledge; MK = Math Knowledge; RC = Reading Comprehension; PS = Physical Science; TR = Table Reading; IC = Instrument Comprehension; BC = Block Counting; AI = Aviation Information; EM = Electrical Maze; GS = General Science; HF = Hidden Figures; DI = Data Interpretation; RB = Rotated Blocks; SR = Scale Reading; *g* = general mental ability; *s* = special cognitive ability. *Based on results of a psychometric evaluation of candidate cognitive subtests for AFOQT Form U (see Walsh, Wilson, et al. in press). **Based on the AFOQT Form T subtest-level analyses (see Walsh et al., 2022). ***The test length and administration times are based on the paper-and-pencil AFOQT forms O, P, Q, R/S, and T Version 1. These values may change with the deployment of the computer-administered Form T Version 2.

As explained in Walsh, Wilson, et al. (in press), the mapping between some of the operational subtests and core officer competencies was weak. Therefore, it was further recommended that the AFOQT be augmented with measures of fluid intelligence and/or specific cognitive abilities. This report is concerned with evaluating candidate DoD batteries and subtests that could bridge this gap.

2.2 DoD Batteries and Subtests Under Consideration

Table 2 summarizes the information that was gathered for each of the DoD batteries and their respective subtests that were considered in this effort. Particular attention was given to the subtests that purported to assess cognitive abilities, which did not require specialized equipment, and that

could easily be adapted to the AFOQT testing format. Therefore, of the large number of eligible subtests, 16 were considered.

Table 2. Summary of the 16 Candidate DoD Batteries and Subtests

Batteries (in alpha order)	Description (in bold are the subtests under consideration)	Officer or Enlisted	Commissioning or Classification	Data Availability
Not Applicable	The **Weight Perception test (WPt)** is based on the Wechsler Adult Intelligence Scales (WAIS-IV) Figure Weights (FW) test. FW was developed to measure fluid reasoning, and more specifically, quantitative and analogical reasoning (Wechsler, 2008). The subtest does not use traditional mathematical content, but instead uses the concept of balancing weights on two sides of a scale. The subtest also involves working memory, as indicated by results of factor analyses (Wechsler, 2008). The involvement of working memory increases with item difficulty, as more difficult items require that a greater number of shape-weight relationships be retained and evaluated to find the correct solution. This relationship between reasoning and working memory is not surprising, based on related research suggesting a dynamic interplay between fluid reasoning, working memory, and processing speed (Kyllonen & Christal, 1990; de Ribaupierre & Lecerf, 2006; Unsworth & Engle, 2007).	Officer	Both	Yes
Aviation Selection Test Battery (ASTB)	The ASTB is the primary test battery used for selecting student pilots and flight officers into aviation training for the US Navy, US Marine Corps, and US Coast Guard. The current version of the ASTB is partially computer-adaptive, and consists of cognitive, non-cognitive, and psychomotor subtests. Cognitive subtests include **Math Skills (MST), Mechanical Comprehension (MCT), Reading Comprehension (RCT), and Aviation and Nautical Information (ANIT).** Non-cognitive subtests include the Naval Aviation Trait Facet Inventory (NATFI) and Biographical Inventory with response verification. The psychomotor portion of the battery is called the Performance-Based Measures (PBM).	Officer	Both	No
Armed Services Vocational Selection	The ASVAB **Assembling Objects (AO)** subtest is a visual and spatial reasoning subtest used to classify recruits into mechanical and engineering jobs (Held et al., 2014). Examinees are required to mentally manipulate objects so that points on the object connect based on a referent picture. The test paradigm requires 16 minutes for the	Enlisted	*Not Applicable*	No

Battery (ASVAB)	paper-and-pencil version and 15 minutes for the computer version. AO is an indicator of fluid intelligence (Held & Carretta, 2013) and contributes both to the general ability and spatial ability factors (Wolfe et al., 1995). The ASVAB **Mental Counters Test (MCt)** is an experimental test of non-verbal reasoning and working memory requiring examinees to add and subtract values that appear on screen for a short time. The test was born out of the Enhanced Computer Administered Test (ECAT) project (Alderton et al., 1997; Larson & Saccuzzo, 1989) from 1990-1992 (Wolf et al., 1995), which aimed to find new tests that could increment validity on the ASVAB. Factor analyses showed MCt to contribute to both the general ability factor and to the working memory factor and is viewed as an indicator of fluid intelligence (Held & Carretta, 2013; Wolfe et al., 1995).			
Electronic Data Processing Test (EDPT)	The EDPT is designed to classify enlisted airmen into rated career fields. The battery consists of four subtests: **Verbal Analogies (VA), Arithmetic Reasoning (AR), Number Series (NS), and Non-Verbal Reasoning (NVR).** NS was designed as a perceptual speed test and NVR was designed as a measure of fluid intelligence that does not rely on verbal knowledge.	Enlisted	*Not Applicable*	Yes
Selection of Unmanned Aerial Systems Personnel (SUPer) battery	The SUPer battery was designed for the selection and classification of the US Navy Unmanned Aerial Systems (UAS) operators (Ackerman, 2018). The battery consists of five cognitive subtests: **Necessary Facts (NECFACT)** measuring quantitative ability; **Memory for Landmarks (MEML)** measuring working memory; **Dial Reading (DRT)** and **Flight Checking (FLTCHK)** both measuring perceptual speed; **Compass Directions (CD)** measuring spatial ability. UAS performance is recorded during Initial Flight Training (IFT) and Instrument Qualification (RIQ; Carretta et al., 2016).	Enlisted	Classification	Yes

3.0 METHOD

The previous section briefly described the cognitive subtests under consideration for inclusion in Form U. The following paragraphs describe the technical approach for evaluating the subtests and additional work that was performed for this effort.

3.1 Technical Approach

Available archival data were analyzed for the candidate subtests using Classical Test Theory (CTT; see Nunnally & Bernstein, 1994) and Item Response Theory (IRT; see Lord & Novick, 2008), when appropriate. As Table 2 shows, data were available for WPt, EDPT, and SUPer. No data were available for the ASTB, AO, and MCt. Consequently, we reviewed the relevant literature for these tests.

Additionally, archival data were analyzed for an experimental test in the AFOQT Form T, PS. PS was added to the AFOQT Form T in 2015 as an unscored experimental measure. Its utility to increment prediction above and beyond the operational AFOQT subtests remains unknown. Finally, we reviewed the military and academic literature on the critical attributes/competencies necessary for the USAF rated career fields (CSO, ABM, unmanned aircraft pilots, and manned aircraft pilots). The review provided a valuable roadmap for future revisions of the AFOQT.

4.0 RESULTS

The previous section described the technical approach. The following section details the results of the evaluation for each subtest.

4.1 Weight Perception Test

The data for WPt were collected between April 2018 and May 2018. The data contained demographic information (gender, race, ethnicity, education, and trainee status), item-level data, and subtest-level data for WPt and the AFOQT TR, and AR subtests. The choice of the subtests included in this data collection is unclear, but perhaps it can be assumed that AR was included as a verbally laden analog to WPt and TR was included as a discriminant measure. The WPt data included 10 items plus the scoring key. The AR data included 15 items plus the scoring key. The WPt, AR, and TR subtest scores were sums of the dichotomously scored items. The initial sample size was 379, dropping to 373 after removing noise in the data.

The sample was relatively equally split by gender (51% females). Most participants were White (68%) and Non-Hispanic (77%). Ninety seven percent of the sample were Basic Military Trainees. Attachment 1 contains the entire output; the tables and paragraphs below contain only key findings.

The WPt item-level difficulty parameters (*b* parameter), expressed as *p*-values (percent correct), ranged from moderate to low, indicating that most items were relatively difficult. No items exhibited ceiling or floor effect. The discriminability parameters, expressed as item-total correlations (ITC), were moderate, indicating relatively good internal consistency among the

items. Most of the items exhibited low mean score subgroup differences (SGDs), with the largest effect sizes observed for Black and White subgroups. This is not surprising given a common finding that cognitive subtests result in moderate to large effect sizes for gender and racial subgroups (Ployhart & Holtz, 2008). As shown in Table 3, WPt on average seemed to be more difficult, discriminable, and reliable compared to AR and TR. Gender mean score SGDs were the highest for WPt and race and ethnicity mean score subgroup differences were the lowest. Notably, on average WPt resulted in lower mean score SGDs than AR. Recall that the two subtests purport to measure quantitative ability, with AR being verbally laden.

Table 3. Subtest-Level Psychometrics Results for WPt, AR, and TR

Subtests	# of Items	# of Response Options	Descriptives				Difficulty			Discriminability			Cronbach's Alpha	Cohen's *d*			
			M	*SD*	Skewness	Kurt.	Min	Max	*M*	Min	Max	*M*		F/M	B/W	H/nH	Avg
WPt	10	5	4.22	3.11	.46	-.99	.25	.56	.42	.46	.67	.54	.84	.49**	.52**	.20	.40
AR	15	5	6.53	3.02	.69	-.11	.17	.84	.46	.17	.51	.32	.72	.40**	.59**	.39**	.46
TR[1]	40	5	21.50	5.72	.09	.09	.45	.75	.58	-.08	.41	.23	.75	.05	.76**	.34**	.38
GPA	N/A	N/A	3.19	.41	.27	-1.02	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.32	.48	.04	.28

Note. Cohen's *d* less than .40 are considered low effect sizes, between .40-.79 moderate effect sizes, and above .80 large effect sizes. WPt = Weight Perception test; AR = Arithmetic Reasoning; TR = Table Reading; GPA = High School Grade Point Average; *M* = Mean; *SD* = Standard Deviation; Kurt. = Kurtosis; Min = Minimum; Max = Maximum; F/M = Female/Male comparison; B/W = Black/White comparison; H/nH – Hispanic/Non-Hispanic comparison; Avg = Average; N/A = Not Applicable; Cronbach's Alpha (see Cronbach, 1951). $^{**}p \leq .01$.

[1]TR in this dataset was compared to TR in the operational AFOQT data. The psychometrics for the two subtests were not comparable, casting doubt on the quality of the current data.

Related to the above, the correlations displayed in Table 4 indicate construct validity with correlations being higher between WPt and AR (both measure similar constructs) than between WPt and TR (both measure different constructs). Although the only criterion in the data was high school Grade Point Average (GPA), it correlated positively, albeit non-significantly, with WPt, suggesting preliminary criterion-related validity.

Table 4. Correlations among WPt, AR, and TR

Subtests	1	2	3	4
1. WPt				
2. AR	.52**			
3. TR	.31**	.35**		
4. GPA	.26	.20	.03	

Note. WPt = Weight Perception test; AR = Arithmetic Reasoning; TR = Table Reading; GPA = High School Grade Point Average. **$p \leq .01$.

In sum, WPt showed preliminary construct validity, some evidence for criterion-related validity, and lower mean score subgroup differences than AR. Additional research is needed to establish criterion-related validity of WPt and to establish its incremental validity relative to the rest of the AFOQT subtests. Based on this information, we recommend the inclusion of the WPt in the AFOQT Form U on an experimental basis.

4.2 Aviation Selection Test Battery

Because no archival data were available, we leveraged the published research to evaluate the ASTB and its subtests. Math Skills Test (MST), Reading Comprehension Test (RCT), Mechanical Comprehension (MCT), and Aviation and Nautical Information Test (ANIT) are measures of crystallized intelligence and are therefore *g*-loaded. Not surprisingly, the ASTB composites have been shown to predict student aviator performance and attrition through primary flight training. With that said, some ASTB subtests are analogous to the AFOQT subtests (i.e., MST = AR and MK; RCT = RC; MCT = MC [no longer in operational use]). There is no mean score SGD reduction evidence for these analogs. In fact, the ASTB composites favor men (Cohen's d^i ranged from .36 to .62) and White test-takers (Cohens' *d* range from .27 to 1.05). Based on this information, we do not recommend the ASTB cognitive subtests for the inclusion in the AFOQT Form U. See Keiser et al. (2020) for more information about the ASTB.

4.3 Assembling Objects

Similar to the ASTB, no archival data were available for AO, therefore we used published research to evaluate it. Factor analyses show that AO contributes to *g* and to a spatial factor (Wolfe et al., 1995). Recent studies show that when experimental ASVAB composites that included AO were compared against operational composites that excluded AO, the experimental composites with AO reduced mean score SGDs for the gender, racial, and ethnic subgroups (Beatty et al., 2020). Mean score SGDs are lower for AO compared to the ASVAB technical knowledge ASVAB subtests,

such as Auto/Shop (Held et al., 2015). On the other hand, experimental ASVAB composites with AO show decrements, albeit minimal, to predictive validity within Air Force Specialty Codes (AFSC; Beatty et al., 2020). Earlier studies demonstrated AO's ability to increment validity (ranging between 1% and 3%) across a wide range of criteria and Military Occupational Specialties (MOS; Carey, 1992; Sager et al., 1997). Based on this information, we recommend the inclusion of AO in the AFOQT Form U on an experimental basis.

4.4 Mental Counters Test

Archival data were not available for MCt; therefore, we reviewed available literature. Factor analyses show that MCt contributes to *g* and to the working memory factor (Wolfe et al., 1995). Recent studies by the US Navy show that when experimental ASVAB composites with MCt are compared against operational composites without MCt, the experimental composites with MCt increase aperture/qualification rates for gender, racial, and ethnic subgroups (Foroughi, 2021). Experimental composites with MCt also show predictive validity, comparable to operational ASVAB composites for a sample of targeted jobs, such as the USAF Air Traffic Controllers (ATC) and the US Navy's Operations Specialists (OS; Foroughi, 2021; Held & Wolfe, 1997; Held & Carretta, 2013). Based on this information, we recommend the inclusion of MCt in the AFOQT on an experimental basis.

4.5 Electronic Data Processing Test

The data for the EDPT included demographic information (gender, race, ethnicity, and accession source) and item- and subtest-level data for the four subtests (Verbal Analogies [VA], Arithmetic Reasoning [AR], Number Series [NS], and Non-Verbal Reasoning [NVR]). Each subtest contained 30 items. The sample size was 509. After removing noise in the data, the final sample size was 499.

The sample was predominantly male (85%), White (63%), Non-Hispanic (82%), with most participants being on active duty (64%). The analyses were performed at the item- and subtest-level using CTT. Attachment 2 contains the entire output; the tables and paragraphs below contain only key findings.

At the item level, the items had a wide range of *p*-values (between .02 to .91), suggesting a broad range of item difficulties across the four subtests. A notable number of items near the end of the test exhibited low *p*-values. It could be a sign that the items got more difficult toward the end. However, an alternative (and more likely) explanation is that most test-takers were not able to get to the end of the test due to lack of time. In fact, the omission rates toward the end of the test ranged between 20% and 30%, which supports that explanation. In terms of discriminability, ITC ranged greatly from -.10 to .58. However, values tended to primarily group between .20 to .50. Mean ITCs were .20 (VA), .24 (AR), and .34 (NS and NVR). These ITCs indicate low to moderate internal consistency. In terms of mean score SGDs, most effect sizes were low, a few were moderate, and none were large. When comparing the magnitudes of the effect sizes between subtests, VA and NVR showed the least amount of moderate effect sizes across gender, racial, and ethnic subgroups.

As shown in Table 5, AR was the most difficult EDPT subtest, followed by VA, NVR, and NS. NVR and NS were the most discriminable subtests followed by AR and VA. All subtests but VA had acceptable Cronbach alphas. In terms of mean score SGD, on average, VA exhibited small effect sizes; AR and NVR exhibited mostly small and some moderate effect sizes; and NS exhibited an equal split of the small and moderate effect sizes.

As shown in Table 6, the correlations among the four EDPT subtests are large enough to suggest that the subtests measured constructs that are related to one another (likely due to *g*) but are not redundant. Overall, VA and AR are already included in the AFOQT. NS was the most difficult subtest and produced the highest mean score subgroup differences. NVR had acceptable psychometrics and resulted in lower mean score subgroup differences than NS. Based on this information, we recommend the inclusion of NVR in the AFOQT Form U on an experimental basis.

Table 5. Subtest-Level Psychometrics of the EDPT Subtests

Subtests	# of Items	# of Response Options	Descriptives				Difficulty			Discriminability			Cron. Alpha	Cohen's *d*				
			M	*SD*	Skew.	Kurt.	Min	Max	*M*	Min	Max	*M*		F/M	B/W	A/W	WnH/WH	Avg
VA	30	5	1.36	.13	.51	.33	.02	.72	.36	-.03	.41	.20	.65	.20	.29*	.21	.37	.27
AR	30	5	1.31	.14	.96	1.25	.08	.59	.30	-.05	.42	.24	.73	.24	.43**	.04	.15	.22
NVR	30	5	1.50	.19	.00	-.44	.16	.85	.49	.11	.49	.34	.83	.43**	.13	.30	.29	.29
NS	30	5	1.60	.18	-.41	-.66	.18	.91	.62	-.10	.58	.34	.83	.63**	.41**	.05	.29	.35

Note. Cohen's *d* less than .40 are considered low effect sizes, between .40-.79 moderate effect sizes, and above .80 large effect sizes. VA = Verbal Analogies; AR = Arithmetic Reasoning; NVR = Non-Verbal Reasoning; NS = Number Series; *M* = Mean; *SD* = Standard Deviation; Skew. = Skewness; Kurt. = Kurtosis; Min = Minimum; Max = Maximum; Cron. Alpha = Cronbach's Alpha; F/M = Female/Male comparison; B/W = Black/White comparison; A/W = Asian/White comparison; WnH/WH = White Non-Hispanic/White Hispanic Comparison; Avg = Average. *$p \leq .05$; **$p \leq .01$.

Table 6. Correlations among the EDPT Subtests

Subtests	1	2	3	4
1. VA				
2. AR	.35**			
3. NVR	.41**	.52**		
4. NS	.46**	.44**	.43**	

Note. VA = Verbal Analogies; AR = Arithmetic Reasoning; NVR = Non-Verbal Reasoning; NS = Number Series. **$p \leq .01$.

4.6 Selection of UAS Personnel

The data for the SUPer battery contained demographic information (gender, race, ethnicity, education, and trainee status), subtest-level data for predictors (Compass Directions [CD], Flight Checking [FLTCHK], Dial Reading [DRT], Necessary Facts [NECFACT], and Memory for Landmarks [MEML]), and multiple criteria data for Initial Flight Training (IFT) and Instrument Qualification (RIQ). The acronyms were adapted from Ackerman (2018). IFT criteria included Merit Assignment Selection System (MASS) Total, Raw Check, Raw Daily, and Raw Academic. The RIQ criteria included MASS Total, Raw Check, Raw Daily, Raw Academic, and Raw Flight commander's ratings. The initial sample was 322. There were 283 total cases with IFT scores and 242 with RIQ scores. After removing careless cases, the total sample varied from $N = 234$ to 283 depending on the criterion.

The sample was predominantly male (91%), White (81%), and non-Hispanic (86%). The vast majority of the participants were UAS trainees (82%). Note that item-level data were unavailable for the predictors and criteria. The analyses were performed at the subtest-level using CTT. Also, each subtest was administered twice, with the plan to create two parallel forms, therefore the results are sometimes presented for the two parallel forms separately (Parts 1 and 2) and together (Overall). Attachment 3 contains the entire output; the tables and paragraphs below contain only key findings.

As shown in Table 7, examinees performed slightly better than chance on CD. Chance responding on a 24-item test with 5 response options would have been 4.8. This may be due to measurement error, low motivation, or other factors. FLTCHK had a rather small mean. This is likely due to the fact that it was designed as a speeded test. NECFACT also resulted in a slightly better than chance sample mean, which is suspect. Chance responding on a 10-item test with 5 or 6 response options would have been 2 or 1.67, respectively. Unfortunately, because we did not have item-level data, we could not examine the omission rates, carelessness rates (a lack of motivation when responding to surveys and thereby may introduce noise to the data; DeSimone et al., 2018), and other factors to determine why the psychometrics turned out this way. Regarding mean score SGDs, all subtests showed small effect sizes on average, except DRT which showed moderate effect sizes.

Table 7. Subtest-Level Psychometrics of the SUPer Subtests

Subtests	# of Items	# of Response Options	Descriptives				Cohen's *d*				
			M	*SD*	Skewness	Kurt.	F/M	B/W	A/W	H/nH	Avg
CD Part 1	24	5	6.00	2.97	.30	-.19	.03	.34	.12	.06	.14
CD Part 2	24	5	7.32	3.37	-.04	-.51	.08	.76**	.15	.06	.26
CD Overall	48	5	6.67	2.85	.08	-.31	.07	.63*	.15	.00	.21
FLTCHK Part 1	150	N/A	49.26	15.04	1.02	6.37	.25	.09	.10	.04	.12
FLTCHK Part 2	150	N/A	54.24	13.99	.83	7.24	.12	.13	.31	.05	.15
FLTCHK Overall	300	N/A	51.75	14.05	.96	7.67	.19	.02	.20	.05	.12
DRT Part 1	78	5	30.42	8.22	-.43	1.18	.32	.36	.66**	.25	.40
DRT Part 2	78	5	29.39	8.46	-.28	.94	.42*	.54*	.59*	.44**	.50
DRT Overall	156	5	29.90	7.80	-.51	1.44	.39	.49	.68**	.39*	.49
NECFACT Part 1	10	5 or 6	4.36	1.69	-.18	-.51	.22	.49	.22	.06	.25
NECFACT Part 2	10	5 or 6	2.72	1.52	.30	-.20	.27	.45	.37	.14	.31
NECFACT Overall	20	5 or 6	3.54	1.30	.06	-.24	.31	.60*	.37	.12	.35
MEML	30	2	24.60	3.22	-1.55	5.44	.20	.34	.02	.21	.19

Note. CD = Compass Directions test; FLTCHK = Flight Check test; DRT = Dial Reading test; NECFACT = Necessary Facts test; MEML = Memory for Landmarks test; *M* = Mean; *SD* = Standard Deviation; Kurt. = Kurtosis; F/M = Female/Male comparison; B/W = Black/White comparison; H/nH – Hispanic/Non-Hispanic comparison; Avg = Average. *$p \leq .05$; **$p \leq .01$.

The correlations among the SUPer subtests indicated that most of the subtests were positively correlated with one another. The strongest correlation was between CD and DRT ($r = .41$, $p \leq .01$; see Table 8).

Table 8 Correlations among the SUPer Subtests

Subtests	Part	CD			FLTCHK			DRT			NECFACT			MEML
		Part 1	Part 2	Overall	Part 1	Part 2	Overall	Part 1	Part 2	Overall	Part 1	Part 2	Overall	
CD	Part 1													
	Part 2	.62**												
	Overall	.89**	.91**											
FLTCHK	Part 1	.29**	.23**	.29**										
	Part 2	.32**	.26**	.32**	.87**									
	Overall	.31**	.26**	.31**	.97**	.97**								
DRT	Part 1	.41**	.45**	.48**	.37**	.42**	.41**							
	Part 2	.38**	.39**	.43**	.34**	.40**	.38**	.75**						
	Overall	.42**	.45**	.48**	.38**	.44**	.42**	.93**	.94**					
NECFACT	Part 1	.33**	.36**	.39**	.07	.12*	.10	.34**	.32**	.35**				
	Part 2	.27**	.29**	.31**	.07	.11*	.09	.34**	.36**	.38**	.30**			
	Overall	.38**	.41**	.44**	.09	.15**	.12*	.42**	.42**	.45**	.83**	.78**		
MEML		.17**	.25**	.24**	.03	.14*	.09	.34**	.36**	.38**	.20**	.15**	.22**	

Note. CD = Compass Directions test; FLTCHK = Flight Check test; DRT = Dial Reading test; NECFACT = Necessary Facts test; MEML = Memory for Landmarks test. *$p \leq .05$; **$p \leq .01$.

The correlations among the SUPer subtests and criteria are presented in Table 9. Generally, CD, DRT, and NECFACT were positively correlated with several criteria. In particular, MASS Total and Raw Daily had the strongest correlations, ranging from .17 to .27. In contrast, FLTCHK and MEML had weak and variable correlations with the criteria. Of note, FLTCHK was negatively correlated with most criteria.

Lastly, stepwise regression was conducted with each criterion as a separate outcome variable and the collection of the SUPer subtests as the predictors. Due to the moderate effect sizes observed for DRT, the first round of regressions was exploratory to examine the incremental validity of DRT above and beyond the rest of the subtests. Although zero-order correlations indicated that DRT was significantly related to multiple criteria, it provided trivial incremental validity above and beyond the other four subtests. Because of the moderate effect sizes and little incremental validity, it was omitted from the subsequent regression models.

As shown in Table 10, CD and NECFACT provided incremental validity across the criteria. The remaining subtests provided no additional explained variance above CD and NECFACT. As such, the optimal models identified by stepwise regression included only CD and NECFACT. Together, CD and NECFACT were significant and positive predictors of MASS Total and Raw Daily scores whereas CD was the only significant predictor for Raw Check, Raw Academic, and Raw Flight CC criteria.

Table 9. Correlations among the SUPer Subtests and Criteria

Subtests	IFT				RIQ				
	MASS Total	Raw Check	Raw Daily	Raw Academic	MASS Total	Raw Check	Raw Daily	Raw Academic	Raw Flight CC
CD	.17**	.08	.21**	.12*	.19**	.19**	.18**	.12	.01
FLTCHK	-.05	-.02	-.08	-.05	-.02	.16*	.00	.02	-.06
DRT	.19**	.06	.27**	.10	.19**	.08	.22**	.14*	-.10
NECFACT	.18**	-.01	.23**	.14*	.20**	.02	.22**	.14*	-.03
MEML	.03	.02	.09	.00	-.03	.02	.06	.00	-.11

Note. CD = Compass Directions; FLTCHK = Flight Check; DRT = Dial Reading; NECFACT = Necessary Facts; MEML = Memory for Landmarks; IFT = Initial Flight Training; MASS = Merit Assignment Selection System; RIQ = Instrument Qualification. *$p \leq .05$. **$p \leq .01$.

Table 10. Stepwise Regression Models for the two SUPer Subtests

Subtests	IFT							
	MASS Total		Raw Check		Raw Daily		Raw Academic	
	β	R^2	β	R^2	β	R^2	β	R^2
CD	.12*	.03	.09	.01	.16*	.05	.09	.01
NECFACT	.15*	.03	-.03	<.001	.18**	.05	.20*	.05
TOTAL		.05		.01		.08		.06

Subtests	RIQ									
	MASS Total		Raw Check		Raw Daily		Raw Academic		Raw Flight CC	
	β	R^2	β	R^2	β	R^2	β	R^2	β	R^2
CD	.15*	.04	.20**	.04	.13*	.03	.08	.01	.14*	.03
NECFACT	.16*	.04	-.04	<.001	.18*	.05	.12	.02	.12	.02
TOTAL		.06		.04		.06		.03		.04

Note. CD = Compass Direction; NECFACT = Necessary Facts; IFT = Initial Flight Training; RIQ = Instrument Qualification; MASS = Merit Assignment Selection System. β is the standardized weight when both predictors are entered in the together; R^2 for each predictor is the amount of variance explained when that predictor is entered individually. Total is the amount of variance explained when both predictors are entered in the model. *$p \leq .05$; **$p \leq .01$.

In sum, the SUPer subtests exhibited mixed psychometrics. Judging by the mean score SGDs, multivariate correlations, and stepwise regressions, we initially recommended the inclusion of CD and NECFACT in the AFOQT Form U on an experimental basis. However, after we made this recommendation, we spoke with Dr. Phillip Ackerman and he discouraged us from pursuing NECFACT further due to difficulties involved in the development of the test. We subsequently stopped pursing CD as it was susceptible to test taking strategies resulting in test compromise. Therefore, our ultimate recommendation was not to include SUPer subtests in the AFOQT Form U.

4.7 Physical Science

4.7.1 Current Approach

PS was included in the AFOQT Form T in 2015 on an experimental basis, where it replaced General Science (GS). Its utility to increment prediction above and beyond the other cognitive subtests was unknown. We re-examined the psychometric properties of PS (PS was initially examined in Walsh et al., 2022). We also re-examined the limited criterion-related validation, initially explored by Kantrowitz et al. (2022). Finally, we performed a new criterion-related validation study using four different samples.

Sample 1 primarily consisted of OTS trainees. The sample came from four main accession sources: OTS Civilians, OTS Active Duty, Air National Guard (ANG), and AF Reserves. There were also 12 participants from the AF Reserve Officer Training Corps. Data were collected from August 2016 to February 2020 (N = 1,416). All demographic data were collected during the AFOQT administration. This sample was primarily male (80%) and White (67%). The data included a variety of officer training criteria. The best criteria to describe overall success within OTS were Final Course Score, Distinguished Graduate Status, and Top Graduate Status. Two criterion composites were created. The first criterion composite was meant to indicate leadership performance (referred to as the Leadership Criterion). This criterion is a unit-weighted combination of Tactical Graded Leader Position 1 (GLP 1), Leader Reaction Course (LRC), Peer Ranking, and Instructor Ranking. All four criteria are meant to assess leadership abilities. Note that the Operational GLP 2 and the Warrior Expeditionary Leadership Problem Solving (WELPS) scores are also meant to assess leadership ability, but had excessive missingness. Furthermore, GLP 2 and WELPS had lower correlations with the other leadership criteria. The second criterion composite assessed academic success (referred to as the Academic Criterion). This criterion is a unit-weighted combination of Academic Assessment 1 (AA1), Academic Assessment 2 (AA2), Informative Briefing (B1), Advocacy Briefing (B2), Informative Paper (P1) and Advocacy Paper (P2). The OTS criteria are described in Appendix A.

Samples 2, 3, and 4 leveraged primary training data from three rated career fields: Pilots (N = 1,906), CSO (N = 658), and ABM (N = 267). The majority of examinees across all samples were male (73%-92%) and White (69%-78%). The main criteria utilized were the MASS scores. MASS scores are composites that indicate the overall assessment of a trainee's airmanship based on indicators of performance, such as academic grades, check flight scores, daily flight scores, and flight commander ratings. MASS scores range from 0 to 100. For more information on these three samples and criteria, see Woolley et al. (2023).

We examined correlations among the various criteria, the AFOQT subtests, and the composite scores. Hierarchical regressions were used to investigate the incremental validity provided by PS above and beyond the currently used composite score.

4.7.2 Review of Previous Analyses

In previous research, PS showed mixed psychometrics (see Walsh et al., 2022). It had good item difficulty levels, with a few overly easy items, based on the CTT and IRT indices. Subtest discriminability was marginally acceptable. At least half of the items had exceptionally low *a* parameters and several items showed relatively low ITCs (≤ .30). Subtest-level SGDs showed some large differences, most notably for White versus Black test-takers. Also, PS showed acceptable internal consistency and test-retest reliability, and primarily loaded on *g*. See Table 11 for a summary of psychometrics.

Table 11. Subtest-Level Psychometrics of PS

AFOQT T1					AFOQT T2				
Descriptives					Descriptives				
N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
36,068	10.55	3.97	.12	-.76	32,929	11.03	3.88	.07	-.68
Difficulty					Difficulty				
Min	Max	*M*	*p*-values ≥ .80/.75	*b* ≤ -2.00	Min	Max	*M*	*p*-values ≥ .80/.75	*b* ≤ -2.00
.29	.84	.52	2 (10%)	1 (5%)	.29	.83	.54	1 (5%)	1 (5%)
Discriminability					Discriminability				
Min	Max	*M*	ITC ≤ .30	*a* ≤ 1.00	Min	Max	Mean	ITC ≤ .30	*a* ≤ 1.00
.22	.59	.42	4 (20%)	10 (50%)	.25	.54	.41	3 (15%)	12 (60%)
Cohen's *d*					Cohen's *d*				
F/M	B/W	A/W	WnH/WH		F/M	B/W	A/W	WnH/WH	
.63	.84	.11	.28		.63	.85	.09	.24	
Cronbach's Alpha			Factor	Test-Retest Reliability	Cronbach's Alpha			Factor	Test-Retest Reliability
.75			*g*	.84	.74			*g*	.83

Note. AFOQT T1 = AFOQT Form T1 Version 1; AFOQT T2 = AFOQT Form T2 Version 1; *N* = Sample Size; *M* = mean; *SD* = Standard Deviation; Min = Minimum; Max = Maximum; ITC = Item-Total Correlations (CTT index of discriminability); *a* = Cronbach's Alpha; Cohen's *d* = Effect Size; F/M = Female/Male comparison; B/W = Black/White comparison; A/W = Asian/White comparison; WnH/WH = White Non-Hispanic/White Hispanic comparison; *b* = *b* parameter (an IRT index of difficulty); *a* = *a* parameter (an IRT index of discriminability); *p*-value = a CTT index of difficulty; *g* = general intelligence factor.

We also examined previous criterion-related validity results from Kantrowitz et al. (2022) who examined potential new AFOQT composites. The only composite that included PS in the final step of the stepwise regression was the Academic Aptitude composite. The criterion used was Academic Average, which came from a Specialized Undergraduate Pilot Training (SUPT) sample. The Academic Aptitude composite is used as an indicator of officership potential, rather than utilized for specific training programs (i.e., SUPT), therefore we argue that this sample and criterion may not have been the ideal source from which to create an Academic Aptitude composite.

4.7.3 Current Results

As previously noted, we utilized four different samples and evaluated correlations and hierarchical regression to examine the criterion-related validity of PS.

In the OTS sample, PS demonstrated weak correlations with most of the criteria (see Table 12). It did, however, have moderate correlations with AA1 and AA2 (r = .22, r = .23, respectively, p < .01) and Academic Composite score (r = .27, p < .01). Note that the Academic Composite score in conjunction with the Officer Suitability Measure (OSM) are used to provide an indicator of

officership potential. The rest of the subtest scores and composites in Table 12 are provided for comparison.

We examined incremental validity of PS above and beyond the AFOQT Academic Aptitude composite using hierarchical regression. Four criteria were examined: Current Score, OTS Distinguished Graduate Status, Academic Composite Score, and Leadership Composite Score. PS explained small but significant incremental variance in Current Score (0.85%, $p < .001$). No significant incremental variance was found for the other three criteria.

In the pilot training sample, no significant correlations were found between PS and the criteria (see Table 13). Therefore, no further analyses were performed. All of the AFOQT subtest scores and the Pilot and OSM composites are provided for comparison.

Table 12. Correlations in the OTS Sample

Criteria	Subtest Scores and Composites											
	Academic	OSM	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI
Academic Assessment 1	.38**	.01	.32**	.22**	.31**	.21**	.37**	.22**	.14**	.12**	.14**	.12**
Academic Assessment 2	.40**	.01	.32**	.23**	.34**	.20**	.40**	.23**	.09**	.12**	.10**	.11**
Briefing 1 Informative	.10**	.05	.11**	.07**	.10**	-.01	.16**	.03	.09**	.09**	.09**	.07**
Briefing 2 Advocacy	.11**	.04	.13**	.07**	.11	.01	.16**	-.03	.10**	.06*	.05	.01
Tactical Graded Leader Position	.01	.08**	.05	.02	.05	-.06*	.06*	-.07**	.10**	.02	.09**	-.01
Operational Graded Leader Position	.07	.06	.08	.01	.07	-.01	.15**	-.03	.04	.07	.06	-.04
Warrior Expeditionary Leadership Problem Solving - Graded Leadership Position	.06	.03	.05	.04	.05	.00	.10**	-.03	.06*	.12**	.08**	.04
Paper 1 Informative	.12**	.03	.09**	.08**	.10**	.08**	.11**	.06*	.09**	.04	.07**	.03
Paper 2 Advocacy	.14**	.01	.10**	.09**	.08**	.10**	.14**	.06*	.09**	.06*	.08**	.03
Standards and Publications Test 1	.17**	.00	.14**	.13**	.13**	.09**	.14**	.08**	.13**	.08**	.10**	.10**
Standards and Publications Test 2	.05	.03	.06*	.04	.04	.00	.09**	.02	.08**	.05	.09**	.04
Leadership Reaction Course	.02	.08**	.04	.03	.03	-.03	.07**	-.03	.07*	.05	.07**	.05
Peer Ranking	.05	.11**	.08**	.02	.06*	-.04	.11**	-.09**	.10**	.06*	.07*	.01
Instructor Ranking	.06*	.11**	.09**	.03	.09**	-.03	.12**	-.08**	.09**	.04	.07*	.01
Current Score	.24**	.11**	.22**	.15**	.19**	.08**	.29**	.06*	.18**	.13**	.17**	.10**
Final Score	.07*	.04	.07*	.05	.05	.04	.06*	-.02	.03	.01	.06*	-.02
OTS Distinguished Graduate Status	.19**	.04	.15**	.13**	.15**	.08**	.21**	.08**	.08**	.09**	.09**	.03
OTS Top Graduate Status	.08**	.03	.06*	.07*	.05	.06*	.04	.02	.04	.01	.04	.03
Leadership Composite Score	.00	.02	.02	.00	-.02	-.02	.03	-.01	.04	.00	.00	-.03
Academic Composite Score	.47**	.01	.39**	.27**	.39**	.25**	.46**	.27**	.13**	.15**	.14**	.14**

Note. OSM = Officer Suitability Measure; VA = Verbal Analogies; AR = Arithmetic Reasoning; WK = Word Knowledge; M = Math Knowledge; RC = Reading Comprehension; PS = Physical Science; TR = Table Reading, IC = Instrument Comprehension; BC = Block Counting; AI = Aviation Information. *$p \leq .05$, **$p \leq .01$.

Table 13. Correlations in the Pilot Training Sample

Criteria	Subtest Scores and Composites											
	Pilot	OSM	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI
CombMASSTotal	.15**	.02	-.04	.02	-.04	-.02	.00	.00	.05	.18**	.05	.04
Cumbrance	.10**	.00	.03	.04	.02	-.02	.04	-.01	.03	.07*	.01	.12**
CombRawDaily	.16**	.06	-.02	.05	-.04	-.01	.05	.03	.09**	.12**	-.02	.18**
CombRawAcad	.04	.06*	-.04	.01	-.05	-.04	.02	-.01	.01	-.01	-.06*	.09**
CombRawFltCC	.03	.07	-.04	.01	-.05	-.04	.02	-.01	.03	-.04	-.10**	.10**

Note. CombMASSTotal = Pilots Merit Assignment Selection System Total for Primary Training; CombRawCheck = Pilots Combined Raw Flight Check Score for Primary Training; CombRawDaily = Pilots Combined Raw Daily Score for Primary Training; CombRawAcad = Pilots Combined Raw Academic Score for Primary Training; CombRawFltCC = Pilots Raw Flight Check; OSM = Officer Suitability Measure; VA = Verbal Analogies; AR = Arithmetic Reasoning; WK = Word Knowledge; M = Math Knowledge; RC = Reading Comprehension; PS = Physical Science; TR = Table Reading, IC = Instrument Comprehension; BC = Block Counting; AI = Aviation Information. *$p \leq .05$, **$p \leq .01$.

In the CSO training sample, PS correlated moderately with CSO Merit Assignment Selection System Total for Primary Training (CSOMASSTotal), CombMASSTotal, and CombRawDaily ($r = 18$, $r = 22$, $r = 13$, respectively, $p < .01$; see Table 14). The rest of the subtest scores and composites are provided for comparison.

Table 14. Correlations in the CSO Training Sample

Criteria	Subtest Scores and Composites											
	CSO	OSM	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI
CSOMASSTotal	.21**	-.04	.13**	.22**	.06	.22**	.13**	.18**	.13**	.24**	.19**	.15**
CombMASSTotal	.17**	-.07	.10*	.17**	.03	.22**	.17**	.22**	.13**	.29**	.18**	.21**
CombRawCheck	.05	.03	.00	.09*	-.04	.00	.05	.07	.12**	.14**	.12**	.12**
CombRawDaily	.15**	-.11**	.03	.14**	.01	.12**	.13**	.13**	.13**	.29**	.19**	.24**
CombRawAcad	-.02	-.01	-.01	-.01	-.01	.02	-.01	.03	-.02	-.01	.00	.00

Note. CSOMASSTotal = CSO Merit Assignment Selection System Total for Primary Training; CombMASSTotal = CSO Combined Merit Assignment Selection System Total for Primary Training; CombRawCheck = CSO Raw Check for Primary Training; CombRawDaily = CSO Combined Raw Daily Score for Primary Training; CombRawAcad = CSO Combined Raw Academic Score for Primary Training; CSO = Combat Systems Officer; OSM = Officer Suitability Measure; VA = verbal Analogies; AR = Arithmetic Reasoning; WK = Word Knowledge; M = Math Knowledge; RC = Reading Comprehension; PS = Physical Science; TR = Table Reading, IC = Instrument Comprehension; BC = Block Counting; AI = Aviation Information. *$p \leq .05$, **$p \leq .01$.

Consequently, we examined incremental validity of PS above and beyond the AFOQT CSO composite for these three criteria. PS provided significant incremental variance for CSOMASSTotal (0.71%, $p < .05$) and CombMASSTotal (1.00%, $p < .01$). No significant incremental variance was found for CombRawDaily.

In the ABM training sample, small correlations were found among PS and ABM Merit Assignment Selection System Total for Primary Training (ABMMASSTotal), CombMASSTotal, and CombRawAcad (r = .13, .12, and .14, respectively, $p < .05$; see Table 15). Hierarchical regressions revealed that PS did not provide significant incremental variance above and beyond the AFOQT ABM composite. The rest of the subtest scores and composites in Table 15 are provided for comparison.

Table 15. Correlations in the ABM Training Sample

Criteria	Subtest Scores and Composites											
	ABM	OSM	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI
ABMMASSTotal	.30**	.04	.17**	.09	.15*	.18**	.18**	.13*	.15*	.15*	.17**	.23**
CombMASSTotal	.28**	.05	.17**	.08	.15*	.17**	.17**	.12*	.15*	.16*	.17**	.22**
CombRawAcad	.26**	.04	.22**	.20**	.14*	.10**	.26*	.14*	.16**	.09	.15*	.17**

Note. ABM = Air Battle Manager; OSM = Officer Suitability Measure; VA = Verbal Analogies; AR = Arithmetic Reasoning; WK = Word Knowledge; M = Math Knowledge; RC = Reading Comprehension; PS = Physical Science; TR = Table Reading, IC = Instrument Comprehension; BC = Block Counting; AI = Aviation Information; ABMMASSTotal = ABM Merit Assignment Selection System Total for Primary Training; CombMASSTotal = ABM Combined Merit Assignment Selection System Total for Primary Training; CombRawAcad = ABM Combined Raw Academic Score for Primary Training. *$p \leq .05$, **$p \leq .01$.

4.7.3 PS Recommendation for Form U

Based on previous research and the current investigation, PS displays some acceptable psychometric properties and some areas for improvement. At the minimum, PS needs to be improved in terms of its item-level discriminability and mean score SGDs, especially for gender. PS shows modest correlations with academic-based criteria for OTS and moderate correlations with CSO and ABM training criteria. However, PS is uncorrelated with Pilot training criteria, and does not provide incremental variance for OTS, or ABM criteria. On a more positive note, PS accounts for up to 1% incremental variance for CSO training criteria. One notable caveat is that this level of incremental variance can be outperformed by non-cognitive measures with lower mean score SGDs. For example, previous research showed that personality increases explained variance between 1.3% and 3.8% beyond the AFOQT CSO composite while reducing mean score SGDs (see Woolley et al., 2023). Given our evaluation, we do not expect that PS would add practical incremental prediction of the outcomes evaluated in this study beyond the other AFOQT subtests. However, additional research is needed to determine if PS may predict outcomes in non-rated career fields such as cyber or intelligence.

4.8 Additional Literature Review

In addition to reviewing the candidate DoD tests for the inclusion in the AFOQT Form U, we also reviewed available literature on the attributes/competencies (referred to as constructs) and their subtests/measures (referred to as assessments) critical for USAF career fields (see Table 16).

Table 16. Literature Review

1. Barelka, A., Barron, L., Coggins, M., Hernandez, S., & Kulpa, P. (2019). *Development and Validation of Air Force Foundational Competency Model*
2. Gonzalez, M., Pena, D. A., Wolliston, D. J., & Haight, N. R. (2019). USAF Assessment of Needs Analysis: Enlisted and Officer AFS Clusters, Task Order #47QFAA18F0043. San Antonio, TX: Operational Technologies Corporation.
3. Kantrowitz, T., Kingry, D., Engelsted, J., Travinin, G., Lovering, E., Gould, M., & PDRI. (2022). *Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: Job Analysis Linkages, Validity, and Subgroup Differences for Current and Alternative Composites* (p. 245). AFRL-RHWP-TR-2022-0038. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
4. Lentz, E., Horgen, K. E., Schneider, R. J., Ferstl, K. L., Kubisiak, U. C., & Borman, W. C. (2009a). *Air Force Officership Survey Volume I: Survey Development and Analyses* (Technical Report). PDRI: Tampa, FL.
5. Lentz, E., Horgen, K. E., Schneider, R. J., Ferstl, K. L., Kubisiak, U. C., & Borman, W.C. (2009b). *Air Force Officership Survey Volume II: Performance requirement linkages and predictor recommendations* (Technical Report). PDRI: Tampa, FL.
6. Shore, C. W., Haight, N., & Martinez, L. (2020*). Identify Potential I/O or Non-I/O Psychology Assessment Tools/Methods: AFOQT Methods to Reduce Adverse Impact*. HQ Air Force Personnel Center.

7. Shore, C. W., Pena, D. A., Gonzalez, M., Haight, N. R., & Wolliston, D. J. (2019). Officer and Enlisted Needs Analysis (AFCAPS-FR-2020-0001). JBSA-Randolph AFB, TX: Strategic Research and Assessment Branch, Air Force Personnel Center.
8. Teachout, M., Shore, C. W., Martinez, L., & Wolliston, D. (2019). *Identifying Potential Measures for Improved Selection and Classification*

Based on the literature review, it became clear that there is a long list of constructs (200+) and a variety of different assessments designed to measure them. Some assessments were (a) developed for enlisted, officers, or both; (b) some were developed for rated (aircrew) career fields or non-rated career fields; (c) some targeted broad and narrow cognitive, non-cognitive, and/or psychomotor attributes; (d) some focused on attributes/competencies that are no longer important, moderately important, or highly important; and (e) some either showed a great deal of promise in reducing mean score SGDs or not at all.

To create a shorter list, we prioritized the constructs that showed initial promise in reducing mean score SGDs and were relatively easy to include in the AFOQT given pragmatic constraints (computer-based administration, administration time, etc.). Based on this, we developed a list of 13 constructs and three assessments. Next, a team of three Industrial/Organizational (I/O) psychologists evaluated whether any of these constructs would improve the AFOQT. First, each psychologist reviewed the literature for a given construct and then shared the findings with the other two psychologists. Finally, they discussed what color code to assign to each construct across the columns. Table 17 summarizes the outcomes of this process. See Attachment 4 and Appendices B – M for more detail.

As shown in Table 17, Based on our evaluation, we recommended that AFPC/DSYX consider the inclusion of subtests that measure inductive reasoning (fluid intelligence), learning/learning agility, short term (working) memory, long term (meaningful) memory, and emotional intelligence. The literature review also revealed perceptual/processing speed as one of the key constructs. However, the AFOQT already includes a direct measure of perceptual/processing speed, called TR. The literature also revealed some soft skills that may need to be assessed; however, we recommended that they be measured by a different instrument than AFOQT, for example a biographical data (biodata) method. Note that these findings are on par with those obtained in the first step of this project described in greater detail in Walsh, Wilson, et al. (in press).

Table 17. Summary of the Literature Review

Constructs (in alpha order)	Incremental Validity*	Mean Score SGD Reduction**	Pragmatic Considerations**	Recommendation
Emotional Intelligence				Consider
Inductive Reasoning (Fluid Intelligence)				Consider
Learning/Learning Agility				Consider
Logic				Do not consider
Long Term (Meaningful) Memory				Consider
Need for Cognition				Do not consider
Perceptual/Processing Speed				Consider
Short Term (Working) Memory				Consider
Soft Skills: Followership				Consider for Biodata
Soft Skills: Growth Mindset				Consider for Biodata
Soft Skills: Instructing				Consider for Biodata
Soft Skills: Moral Courage				Consider for Biodata

*Incremental validity: Green = empirical evidence exists and shows non-trivial incremental validity; Orange = some empirical evidence exists but the results may be mixed; Red = empirical evidence suggests that the incremental validity is trivial; Grey = no empirical evidence exists. **SGD Reduction: Green = empirical evidence exists and shows nontrivial reduction in SGD for all gender, racial, and ethnic subgroups; Orange = some empirical evidence exists but the results may be mixed such that some subgroups may see a reduction but not others; Red = empirical evidence suggests that the reduction is trivial; Grey = no empirical evidence exists. **Pragmatic Considerations: Green = there is an existing commercial or off-the-shelf measure; Orange = there is an existing measure, but it may need to be modified and adapted to the AFOQT populations and purposes; Red = there is no existing measure; a measure would need to be developed from scratch; or a measure would need to be evaluated upon integration.

5.0 DISCUSSION

The effort described in this technical report served as the second step in development of AFOQT Form U. The primary goal of this review was to identify available DoD subtests that would complement the next generation AFOQT. The secondary goal was to review available literature to see what other constructs should be considered.

We reviewed 16 candidate DoD cognitive batteries and subtests. WPt showed preliminary construct validity, some evidence for criterion-related validity, and lower mean score SGDs than AR. Additional research is needed to establish criterion-related validity for WPt and to establish its incremental validity relative to the current AFOQT subtests. Based on this information, we recommend the inclusion of the WPt in the AFOQT Form U on an experimental basis.

The four ASTB cognitive subtests are conceptually and psychometrically similar to the AFOQT subtests. Therefore, we recommended not including them in AFOQT Form U.

AO's research is mixed with some suggesting that it increments prediction while others suggesting it does not. However, there is evidence that operational composites with AO included tend to reduce mean score SGDs. Based on this information, we recommended the inclusion of the AO in AFOQT Form U on an experimental basis.

MCt showed promise in terms of prediction and reduction of mean score SGDs. We recommend the inclusion of the MCt in the AFOQT on an experimental basis.

Of the four EDPT subtests, two are already included in the AFOQT. The two that are not yet included are NS and NVR. NS was the most difficult subtest and produced the highest mean score SGDs. NVR had acceptable psychometrics and resulted in lower mean score SGDs than NS. We recommend the inclusion of NVR in the AFOQT Form U on an experimental basis.

Item-level data were not available for the SUPer subtests. Therefore, the analyses were performed at the subtest level only. In sum, the SUPer subtests exhibited mixed psychometrics. Judging by the mean score SGDs, multivariate correlations, and stepwise regressions, we initially recommended the inclusion of CD and NECFACT in the AFOQT Form U on an experimental basis. However, after we made this recommendation, we spoke with Dr. Phillip Ackerman who discouraged us from pursuing NECFACT due to the difficulties involved in the development of the test content. In subsequent months, we stopped pursing CD for AFOQT and instead began considering it for inclusion as an experimental subtest for the Test of Basic Aviation Skills (TBAS). Our recommendation was not to include SUPer subtests in the AFOQT Form U (but consider CD for TBAS).

Next, we examined criterion related validity for PS across four samples. In the OTS sample, PS showed modest correlations with academic-based criteria and did not provide practical incremental variance over the AFOQT Academic Aptitude composite. In the Pilot sample, PS did not show significant correlations with training criteria. In the CSO sample, PS showed moderate correlations with training criteria and provided up to 1% incremental variance above and beyond the AFOQT CSO composite for some training criteria. In the ABM sample, PS showed moderate correlations with training criteria but did not provide incremental variance above and beyond the AFOQT Form T ABM composite. The incremental prediction may be outweighed with other noncognitive subtests, such as a personality inventory, which may also reduce mean score SGDs. In conclusion, we do not expect that PS would add practical incremental prediction beyond the AFOQT Form T subtests for the prediction of aircrew training outcomes. However, more research is needed to determine its utility for predicting outcomes in non-rated career fields.

The final step in this effort was a review of the available literature about the merits of additional cognitive abilities that may augment future versions of the AFOQT. We recommend that the Air Force policy makers and practitioners consider inclusion measures of inductive reasoning (fluid intelligence), learning/learning agility, short term (working) memory, long term (meaningful) memory, and emotional intelligence. Note that the literature review also revealed perceptual/processing speed as one of the key constructs. However, the AFOQT already includes

a direct measure of perceptual/processing speed, called TR. Assessment of soft skills can on AFOQT Form T is provided by the SJT and SDI-O subtests, These could be augmented by including other non-cognitive measures such as biodata. Reviewing this information against the experimental subtests that we recommended for inclusion in From U, it appears that some of the subtests may capture the missing constructs. Specifically, MCt, AO, NVR, and WPt may capture fluid intelligence, working memory, and perceptual speed.

6.0 CONCLUSION

This effort and others (e.g., Kantrowitz et al., 2023; Walsh, Brady, et al., 2022; Walsh, Wilson, et al., in press; Walsh, Woolley, et al., 2022) laid the foundation for AFOQT Form U development. The Form U blueprint includes nine Form T cognitive subtests (VA, AR, MK, RC, PS, TR, IC, BC, and AI), two non-cognitive subtests (SJT and SDI-O), and four experimental subtests (AO, NVR, WPt, and MCt). The expectation is that the experimental subtests will capture some of the missing constructs (thereby expanding the competencies measured in commissioning officers and officers pursuing rated career fields) and also increase aperture for historically underrepresented gender, racial, and ethnic subgroups. The next step in the project entails generating a large pool of items for each subtest and collecting data for initial psychometric evaluation.

7.0 REFERENCES

Ackerman, P. L. (2018). *Selection and Classification for UAS Personnel (SUPer): Technical Area #1*. Contract N00014-14-C-0051. Office of Naval Research.

Alderton, D. L., Wolfe, J. H., & Larson, G. E. (1997). The ECAT battery. *Military Psychology*, *9*(1), 5-37.

Barelka, A., Barron, L., Coggins, M., Hernandez, S., & Kulpa, P. (2019). *Development and Validation of Air Force Foundational Competency Model.* JBSA Randolph AFB, TX: HQ Air Education and Training Command.

Beatty, A. S., Burke, M. I., Koch, A. J., Trippe, D. M. (2020). Next-generation Armed Services Vocational Aptitude Battery (ASVAB). AFRL-RH-WP-TR-2020-0099. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Carey, N. B. (1992). Does Choice of a criterion matter? *Military Psychology*, *4*(2), 103-117.

Carretta, T. R., Rose, M. R., & Bruskiewicz, K. T. (2016). Selection methods for operators of remotely piloted aircraft systems. In N. J. Cooke, L. Rowe, W. Bennett, & D. Q. Joralmon (Eds.), *Remotely Piloted Aircraft Systems: A Human Systems Integration Perspective: A Human Systems Integration Perspective*, 137-162.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

de Ribaupierre, A., & Lecerf, T. (2006). Relationships between working memory and intelligence from a developmental perspective: Convergent evidence from a neo-Piagetian and a psychometric approach. *European Journal of Cognitive Psychology*, *18*(1), 109-137.

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology*, *67*(2), 309-338.

Foroughi, C. K. (2021). Mental Counters Study of Instructions. Manpower Accession Policy Working Group.

Gonzalez, M., Pena, D. A., Wolliston, D. J., & Haight, N. R. (2019). *USAF Assessment of Needs Analysis: Enlisted and Officer AFS Clusters*, Task Order #47QFAA18F0043. San Antonio, TX: Operational Technologies Corporation.

Held, J. D., & Carretta, T. R. (2013). *Evaluation of tests of processing speed, spatial ability, and working memory for use in military occupational classification* (p. 0035). Navy Personnel Research, Studies, & Technology (NPRST/BUPERS-1).

Held, J. D., & Wolfe, J. H. (1997). Validities of unit-weighted composites of the ASVAB and the

ECAT battery. *Military Psychology*, *9*(1), 77-84.

Held, J. D., Carretta, T. R., & Rumsey, M. G. (2014). Evaluation of tests of perceptual speed/accuracy and spatial ability for use in military occupational classification. *Military Psychology*, *26*(3), 199-220.

Kantrowitz, T., Kingry, D., Engelsted, J., Travinin, G., Lovering, E., & Gould, M. (2022). *Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: Job Analysis Linkages, Validity, and Subgroup Differences for Current and Alternative Composites* (p. 245). AFRL-RHWP-TR-2022-0038. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Kantrowitz, T., Segall, D., Kingry, D., Valone, A., Mann, K., Walsh, J., Wilson, R., Trent, J., & Carretta, t. (2023, June). *Development of the Air Force Officer Qualifying Test (AFOQT) Form T version 2,* AFRL-RH-WP-TR-2023-0021. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Keiser et al. (2020, May). Aviation Selection Test Battery (ASTB) Annual Report. Operational psychology department. Navy Medicine Operational Training Center; Naval Aerospace Medical Institute Detachment; Naval Air Station Pensacola.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*(4), 389-433.

Larson, G. E., & Saccuzzo, D. P. (1989). Cognitive correlates of general intelligence: Toward a process theory of g. *Intelligence*, *13*(1), 5-31.

Lentz, E., Horgen, K. E., Schneider, R. J., Ferstl, K. L., Kubisiak, U. C., & Borman, W. C. (2009a). *Air Force Officership Survey Volume I: Survey Development and Analyses* (Technical Report). PDRI: Tampa, FL.

Lentz, E., Horgen, K. E., Schneider, R. J., Ferstl, K. L., Kubisiak, U. C., & Borman, W.C. (2009b). *Air Force Officership Survey Volume II: Performance requirement linkages and predictor recommendations* (Technical Report). PDRI: Tampa, FL.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. NY: McGraw-Hill.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*(1), 153-172.

Sager, C. E., Peterson, N. G., Oppler, S. H., Rosse, R. L., & Walker, C. B. (1997). An examination of five indexes of test battery performance: Analysis of the ECAT battery. *Military*

Psychology, *9*(1), 97-120.

Shore, C. W., Haight, N., Martinez, L., & HQ Air Force Personnel Center. (2020). *Identify Potential I/O or Non-I/O Psychology Assessment Tools/Methods: AFOQT Methods to Reduce Adverse Impact* (p. 125). unpublished technical report. Randolph AFB. TX Air Force Personnel Center, Strategic Research and Assessment Branch.

Shore, C. W., Pena, D. A., Gonzalez, M., Haight, N. R., & Wolliston, D. J. (2019). *Officer and Enlisted Needs Analysis* (AFCAPS-FR-2020-0001). JBSA-Randolph AFB, TX: Strategic Research and Assessment Branch, Air Force Personnel Center.

Teachout, M., Shore, C. W., Martinez, L., & Wolliston, D. (2019). *Identifying potential measures for improved selection and classification*. San Antonio, TX: Operational Research Group.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological review*, *114*(1), 104.

Walsh, J. L., Brady, M. F., Woolley, M. R., Carretta, T. R., & Infoscitex. (2022). *Air Force Officer Qualifying Test (AFOQT) Form T evaluation: Item-level analyses* (p. 139). AFRL-RH-WP-TR-2022. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Performance Optimization Branch.

Walsh, J. L., Wilson, R., Mann, K.J., Sizemore, S., Trent, J., & Carretta, T. R. (in press). *Air Force Officer Qualifying Test (AFOQT) Form U: Reviewing cognitive subtests from previous forms,* AFRL-RH-WP-TR-2024-xxxx. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Performance Optimization Branch.

Walsh, j. L., Woolley. M. R., Brady, M. F., & Carretta, T. R. (2022). *Air Force Officer Qualifying test (AFOQT) Form T evaluation: Subtest-level analyses,* AFRL-RH-WP-TR-2022-0022. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Performance Optimization Branch.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition technical and interpretive manual*. San Antonio, TX: Pearson.

Wolfe, J. H., Alderton, D. L., Larson, G. E., & Held, J. D. (1995). *Incremental validity of enhanced computer administered testing (ECAT).* San Diego, CA: Navy Personnel Research and Development Center.

Woolley, M., Walsh, J., Carretta, T., Mouton, A., & Deregla, A. (2023). *Development and psychometric evaluation of predictive success models for US Air Force rated career f fields,* AFRL-RH-WP-TR-2023-0007. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Systems Directorate, Airman Biosciences Division, Performance Optimization Branch.
.

Appendix A – OTS criteria described in Section 4.7

Criterion	Definition
Academic Assessment 1	Academic assessments and the Pre-Course Assignment assessment are tests derived from academic coursework. The questions are derived from Samples of Behavior from each lesson. Lessons Covered: Air Force Leader Development, Suicide Prevention, Religious Accommodation, Team Building, Problem Solving, Sexual Assault Prevention and Response, Full Range Leadership, Change Management, Followership, Conflict Management.
Academic Assessment 2	Academic assessments and the Pre-Course Assignment assessment are tests derived from academic coursework. The questions are derived from Samples of Behavior from each lesson. Lessons Covered: Cyberspace, National Security Strategy, Military Justice, Professional and Unprofessional Relationships, Managing in a Diverse World, Counseling and Practicum, Cross-Cultural Competence.
Briefing 1 Informative	This briefing is an oral presentation. Topic guidance will be given to trainees during the briefing description. It will be presented in compliance with military briefing guidelines and standards.
Briefing 2 Advocacy	This briefing is an oral presentation. Topic guidance will be given to trainees during the briefing description. It will be presented in compliance with military briefing guidelines and standards.
Tactical Graded Leader Position	Tactical GLPs are designed to evaluate a trainee's ability to apply the core competencies of officership and lead his or her peers in a field environment. These positions are designed to challenge test a trainee's leadership qualities within a short, constrained scenario.
Operational Graded Leader Position	Operational GLPs are designed to evaluate a trainee's ability to apply the core competencies of officership in a formal leadership role within the typical Air Force wing structure.
Warrior Expeditionary Leadership Problem Solving - Graded Leadership Position	WELPS is a problem solving, scenario-based field exercise designed to allow trainees to apply concepts of leadership, followership, problem solving methods, communication, team building, and motivation techniques in a small group under time constraints. Trainees will also assess their role in the group and how they react to group dynamics. This practice exercise is not formally evaluated but is an opportunity for the trainee to receive feedback regarding strengths/weaknesses in field leadership.
Paper 1 Informative	Operational GLP can also be accomplished by assigning a trainee a significant additional duty for the duration of OTS. These positions are designed to mimic real world leadership responses and demands as observed in baseline leader assignments.

Paper 2 Advocacy	The second paper is a written assignment that provides background information to its reader in a manner provided in the descriptive briefing for the assignment. It will be written in an approved format IAW AFH 33-337 Tongue & Quill (most recent edition) and AU-1 Air University Style and Author Guide (most recent edition).
Standards and Publications Test 1	Trainees are tested on their knowledge of the OTSMAN 36-2604. Minimum passing score is 80 percent.
Standards and Publications Test 2	Trainees are tested on their knowledge of the OTSMAN 36-2604. Minimum passing score is 80 percent.
Leadership Reaction Course	LRC is a course designed to provide trainees the opportunity to display individual leadership and followership traits, lead teams and problem solve in a time-sensitive environment. LRC consists of scenario-based missions where trainees are evaluated on how well they lead their team and their application of academic concepts to the problem-solving environment.
Peer Ranking	This lesson has a dual purpose. First, it gives the trainees a chance to apply the rating process by rating their peers. Secondly, it shows the importance of evaluations as a means of self- improvement. All instructions for trainees to complete the peer evaluations are included in the study guide. Instructors should conduct feedback sessions with individual trainees regarding the information resulting from the peer evaluations.
Instructor Ranking	Flight instructors will conduct a thorough review of trainees to evaluate their mental, moral, physical, and professional fitness. Upon conclusion of the review the flight instructor will recommend graduation or other action. When the flight instructor and the reviewing authorities do not recommend a trainee for graduation, the TRS Sq/CC determines if elimination action is appropriate. Flight instructors give feedback based on this review to individual trainees during the last week of training. This counseling summarizes the overall performance of the trainee throughout OTS.
Current Score	Earlier overall score in OTS program.
Final Score	Final overall score in OTS program.
OTS Distinguished Graduate Status	Student achieved distinguished graduate status.
OTS Top Graduate Status	Student achieved top graduate status.

Leadership Composite Score	This criterion is a unit-weighted combination of Tactical Graded Leader Position (GLP) 1, Leader Reaction Course (LRC), Peer Ranking, and Instructor Ranking. All four criteria are meant to assess leadership abilities. Note that the Operational GLP 2 and the Warrior Expeditionary Leadership Problem Solving (WELPS) scores are also meant to assess leadership ability but had too much missingness in this dataset to be utilized. Furthermore, GLP 2 and WELPS had lower correlations with the other leadership criteria.
Academic Composite Score	This criterion is a unit-weighted combination of Academic Assessment 1 (AA1), Academic Assessment 2 (AA2), Informative Briefing (B1), Advocacy Briefing (B2), Informative Paper (P1) and Advocacy Paper (P2).

Appendix B – Emotional Intelligence

Definition:

- Emotional intelligence (EI) is a conceptual area that involves traits (trait EI) and abilities (ability EI) pertaining to emotional perception, understanding, facilitation, and regulation.
- Trait EI is better measured with a self-report and ability EI is better measured with a maximum performance test (Petrides, 2011).

SGD Reduction Evidence:

- The degree of the reduction in mean score subgroup differences depends on the form of measurement:
  - *Performance-based EI measures* seem to have the highest Black and White mean score subgroup differences (Cohen's d = .99), followed by men-women mean score subgroup differences (Cohen's d = .52, favoring women; Rhodes, 2008).
  - *Self-report EI measures* seem to have relatively low Black-White mean score subgroup differences, favoring Black test-takers (Cohen's d = .21; Rhodes, 2008).
  - *Mixed-report measures* seem to have lower mean score subgroup differences across the board (Rhodes, 2008).
  - *The aggregate EI measures* seem to have low Black-White mean score subgroup differences (Cohen's d = .17; Joseph & Newman, 2010).

Incremental Validity Evidence:

Some studies show that trait emotional intelligence explains additional variance in well-being outcomes above and beyond personality (Petrides et al., 2007). Other studies show that emotional intelligence measures predict deviant behaviors and alcohol use above and beyond measures of personality and verbal intelligence (Brackett et al., 2003).

References

Brackett, M. A., & Mayer, J. D. (2003). Convergent, discriminant, and incremental validity of competing measures of emotional intelligence. *Personality and Social Psychology Bulletin*, *29*(9), 1147-1158.

Joseph, D. L., & Newman, D. A. (2010). Emotional intelligence: an integrative meta-analysis and cascading model. *Journal of applied psychology*, *95*(1), 54.Petrides, K. V. (2011). Ability and trait emotional intelligence. In T. Chamorro-Premuzic, S. von Stumm, & A. Furnham (Eds.), *The Wiley-Blackwell handbook of individual differences* (pp. 656–678). Wiley Blackwell.

Petrides, K. V., Pérez-González, J. C., & Furnham, A. (2007). On the criterion and incremental validity of trait emotional intelligence. *Cognition and emotion*, *21*(1), 26-55.

Rhodes, D. L. (2008). *Is Emotional Intelligence Worthwhile? Assessing Incremental Validity and Adverse Impact.* [Master's Thesis, Texas A & M University].

Appendix C – Inductive Reasoning (Fluid Intelligence)

Definition:

- Roots of inductive and deductive reasoning lie in formal logic.
- Psychologists sometimes define deduction as applying a general principle/rule to a specific case and induction as observing a series of specific cases and inferring/formulating a general principle/rule (Colberg et al. 1985).
- With deductive reasoning, information is complete and conclusions are described as absolutely valid or absolutely invalid. With inductive reasoning, information is incomplete and conclusions have a probability of truth.

SGD Reduction Evidence*:

- There are mean score differences between Black-White subgroups, but they are lower than for standard *g*-loaded measures (Cohen's *d* range from .39 to .66; Hausdorf et al., 2003; McKay et al., 2003). Most of these results were based on outcomes of Raven's Standard Matrices, which could be argued to measure working memory and not reasoning. However, the classification of these tests as fluid intelligence or inductive reasoning is also quite common.
- Some studies report that Black subgroup score about 1 standard deviation below White subgroup (McKay et al., 2002; 2003; Rushton et al., 2002).
  - Some of these effects seemed to be driven by stereotype threat. There is evidence to suggest that under optimal conditions (lowering stereotype threat) the Black-White mean score subgroup differences would mostly disappear (Brown & Day, 2006).

Incremental Validity Evidence:

- Inductive reasoning (as measured by Raven's Progressive Matrices) has been shown to correlate with specific forms of performance, such as managerial performance (.20) and interpersonal skill performance (.12) and is suggested to explain about 10%-25% of variance in general performance across a wide range of occupational settings (Raven, 2000)
  - Not much specific evidence regarding incremental prediction over *g*

References

Brown, R. P., & Day, E. A. (2006). The difference isn't black and white: stereotype threat and the race gap on Raven's Advanced Progressive Matrices. *Journal of Applied Psychology*, *91*(4), 979.

Colberg, M., Nester, M. A., & Trattner, M. H. (1985). Convergence of the inductive and deductive models in the measurement of reasoning abilities. *Journal of Applied Psychology*, *70*(4), 681.

Corsini, R. J. (1957). *Nonverbal Reasoning.* Park Ridge, IL: London, House.

Hausdorf, P. A., LeBlanc, M. M., Chawla, A. (2003). Cognitive ability testing and employment

selection: Does test content relate to adverse impact? *Applied Human Resource Management Research, 7,* 41-48.

Klein, R. M., Dilchert, S., Ones, D. S., & Dages, K. D. (2015). Cognitive predictors and age-based adverse impact among business executives. *Journal of Applied Psychology*, *100*(5), 1497.

McKay, D. R., Ridding, M. C., Thompson, P. D., & Miles, T. S. (2002). Induction of persistent changes in the organization of the human motor cortex. *Experimental Brain Research*, *143*(3), 342-349.

McKay, P. F., Doverspike, D., Bowen-Hilton, D., & McKay, Q. D. (2003). The effects of demographic variables and stereotype threat on black/white differences in cognitive ability test performance. *Journal of Business and Psychology, 18,* 1-14.

Raven, J. (2000). Psychometrics, cognitive ability, and occupational performance. *Review of Psychology, 7,* 51-74.

Rushton, J. P., Skuy, M., Fridjhon, P. (2002). Jensen effects among African, Indian, and White engineering students in South Africa on Raven's Standard Progressive Matrices. *Intelligence, 30,* 409-423.

Vangent. (1993). *Computer Programmer Aptitude Battery examiner's manual.* Chicago, IL: Author.

Appendix D – Learning/Learning Agility

Definition:

- Learning ability (aka learning agility) is the ability to learn from experience and apply that learning to perform successfully in the future (Lombardo & Eichinger, 2004).
- According to this view, high learning agility individuals learn the right lessons from experience and apply those lessons to novel situations (De Meuse et al., 2010).

SGD Reduction Evidence:

- There is a modest correlation between learning ability and *g* (*r* range from .11 to .33; Ogisi, 2006).
- One study reported no significant mean score differences for gender subgroups (Bedford, 2011).
- A different study found that mean score subgroup differences varied by scale, but even then Cohen's *d* were lower compared to Cohen's d found for *g* (Cohen's d ranged from .20 to .30s; De Meuse et al., 2011).
- No significant mean score differences were reported for White-Asian subgroups; the same study also compared White subgroup and "other minorities" which included groups such as Native Americans and also found no significant mean score difference (De Meuse et al., 2011).

Incremental Validity Evidence:

- There is some evidence that learning agility add incremental variance above and beyond measures of personality and cognitive ability for supervisory ratings of both promotability and overall job performance (Allen, 2016; Connolly, 2002).

References

Allen, J. (2016). Conceptualizing learning agility and investigating its nomological network.

Bedford, C. L. (2011). *The Role of Learning Agility in Workplace Performance and Career Advancement (UMI# 3465093.)* [Doctoral dissertation, University of Minnesota]. Proquest.

Connolly, J. (2001, Dissertation). Assessing the construct validity of a measure of learning agility. FIU Electronic Theses and Dissertations. 2424. https://digitalcommons.fiu.edu/etd/2424

De Meuse, K. P., Dai, G., & Hallenbeck, G. S. (2010). Learning agility: A construct whose time has come. *Consulting Psychology Journal: Practice and Research*, *62*(2), 119.

Lombardo, M. M., & Eichinger, R. W. (2004). FYI: For your improvement. *A Guide for Development and Coaching.* Lominger Ltd.

Ogisi, M. (2006). *Assessing learning agility and its relationship to personality, cognitive ability, and learning styles (UMI# 1438522.)* [Master's thesis, Northern Kentucky University]. Proquest.

Appendix E – Logic

Definition:

- Logic is defined as the ability to apply reasoning to problems by relying on strict principles of validity.

SGD Reduction Evidence:

- Logic-based tests seem to yield relatively low Black-White mean score subgroup differences (Cohen's *d* = .38; De Soete et al., 2013; Yusko et al., 2012).
- "Low" *g*-loaded version of a logic-based test yielded a Cohen's *d* of .47 for Black-White subgroups and a Cohen's d of .19 for Hispanic-White subgroups (McDaniel, 2018). Significance tests were not included. It seems the authors simply took a logic-based test and reduced reading requirements (based on the Siena Reasoning Test).

Incremental Validity Evidence:

- Generally, verbally-laden logic-based measures are expected to show little incremental variance above and beyond *g* (Stauffer et al., 1996; Trippe et al., 2014).

References

De Soete, B., Lievens, F., & Druart, C. (2013). Strategies for dealing with the diversity-validity dilemma in personnel selection: Where are we and where should we go? *Revista de Psicología del Trabajo y de las Organizaciones*, *29*(1), 3-12.

Yusko, K. P., Goldstein, H. W., Scherbaum, C. A., & Hanges, P. J. (2012, April). Siena Reasoning Test: Measuring intelligence with reduced adverse impact. In *27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA*.

McDaniel, M. A. (2018). *How to build a cognitive ability test with reduced mean group differences.* [Conference presentation]. Personnel Testing Council of Metropolitan Washington.

Stauffer, J. M., Ree, M. J., & Carretta, T. R. (1996). Cognitive-components tests are not much more than g: An extension of Kyllonen's analyses. *The Journal of General Psychology*, *123*(3), 193-205.

Trippe, D. M., Moriarty, K. O., Russell, T. L., Carretta, T. R., & Beatty, A. S. (2014). Development of a cyber/information technology knowledge test for military enlisted technical training qualification. *Military Psychology*, *26*(3), 182-198.

Appendix F – Long Term (Meaningful) Memory

Definition:

- Long-term memory (LTM) constitutes any form of memory that expands beyond short-term memory (effectively anything that is stored longer than a few seconds).
- Meaningful Memory (MM) is a type of LTM and represents individual differences in ability to learn and recall information that has meaningful relationships.

SGD Reduction Evidence:

- Evidence exists that LTM (particularly MM) has little to no mean score differences between White-Hispanic subgroups (Cohen's d = .07; Cucina et al., 2015).
- The same study provides evidence that MM has little mean score differences between women and men, favoring women (Cohen's d = .37; Cucina et al., 2015).
- However, another study suggests that LTM may differ in which gender it favors with men performing better in visuospatial LTM and women performing better on verbal and auditory LTM (Pauls et al., 2013).
- Some suggestion of age-based differences (Hardy et al., 2007).

Incremental Validity Evidence:

- LTM, specifically meaningful memory has displayed incremental validity over and above g in training performance (Change-R^2 = .03) and increased model fit (which included g, MM and training performance; Cucina et al., 2015).

References

Cucina, J. M., Su, C., Busciglio, H. H., & Peyton, S. T. (2015). Something more than g: Meaningful Memory uniquely predicts training performance. *Intelligence*, *49*, 192-206.

Pauls, F., Petermann, F., Lepach, A. C. (2013). Gender differences in episodic memory and visual working memory including the effects of age. *Memory, 7,* 857-874.

Hardy, D. J., Satz, P., D'Elia, L. F., Uchiyama, C. L. (2007). Age-related group individual differences in aircraft pilot cognition. *The International Journal of Aviation Psychology, 17,* 77-90.

Appendix G – Need for Cognition

Definition:

- Need for Cognition (NC) has been defined as the "tendency to engage in and enjoy effortful cognitive activity" (Lins de Holanda Coelho, 2020, p. 1870).
- "For people high in NC, thinking satisfies a desire and is enjoyable. For people low in NC, thinking can be a chore that is engaged in mostly when some incentive or reason is present." (Petty et al., 2009, p. 318).
- "NC is only moderately related to measures of cognitive ability (e.g., verbal intelligence) and continues to predict relevant outcomes after cognitive ability is controlled (e.g., Cacioppo et al., 1996; Petty et al, 2009).

SGD Reduction Evidence:

- No literature found

Incremental Validity Evidence:

- There is initial evidence that NC provide incremental prediction above and beyond personality and crystallized intelligence (Fleischhauer et al., 2010).

References

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197.

Fleischhauer, M., Enge, S., Brocke, B., Ullrich, J., Strobel, A., & Strobel, A. (2010). Same or different? Clarifying the relationship of need for cognition to personality and intelligence. *Personality and Social Psychology Bulletin*, *36*(1), 82-96.

Lins de Holanda Coelho, G., HP Hanel, P., & J. Wolf, L. (2020). The very efficient assessment of need for cognition: Developing a six-item version. *Assessment*, *27*(8), 1870-1885.

Petty, R. E., Briñol, P., Loersch, C., & McCaslin, M. J. (2009). *The need for cognition*.

Appendix H – Perceptual/Processing Speed

Definition:

- **Perceptual speed** is defined as the ability to rapidly recognize and compare specific images, numbers, letters or other sensory stimuli.

SGD Reduction Evidence:

- Women generally score higher than men in primary perceptual speed areas (e.g., visual matching tasks, table reading). However, one test did show that in a specific sample (officer candidates), men scored slightly higher than women (Cohen's *d* = .11; Barron & Rose, 2013).
- One study showed that the majority subgroup scored higher than the minority subgroups (White-Black Cohen's d = .78; White-Hispanic Cohen's *d* = .26; Barron & Rose, 2013).
- The Cohen's *d* magnitudes were slightly lower for pilot candidates.

Incremental Validity Evidence:

- Some studies showed incremental validity for measures of perceptual speed/accuracy above and beyond GMA in prediction of task performance (Mount et al., 2008).
- Other studies showed that perceptual speed has incremental validity above academic and technical aptitude for predicting pilot trainee flying performance (Johnson et al., 2017).

Definition:

- **Processing speed** is defined as the ability to rapidly perform simple or automatic cognitive tests. Note that processing speed and perceptual speed are closely related.

SGD Reduction Evidence:

- Gender mean score subgroup differences seem to favor women in major areas (digit/symbol-based tasks). For example, one study reported Cohen's *d* of .46 for the digit symbol test of the WAIS-III (Irwing, 2012).
- Men performed better in some niche tests, such as reaction speed and assembling objects (Irwing, 2012; Roivainen, 2011).

Incremental Validity Evidence:

- See Perceptual Speed

References

Barron, L. G., & Rose, M. R. (2013). Relative validity of distinct special abilities: An example with implications for diversity. *International Journal of Selection and Assessment, 21,* 400-406.

Irwing, P. (2012). Sex differences in *g*: An analysis of the US standardization sample of the WAIS-III. *Personality and Individual Differences, 53,* 126-131.

Roivainen, E. (2011). Gender differences in processing speed: A review of recent research. *Learning and Individual Differences, 21,* 145-149.

Johnson, J. F., Barron, L. G., Carretta, T. R., & Rose, M. R. (2017). Predictive validity of spatial ability and perceptual speed tests for aviator training. *The International Journal of Aerospace Psychology*, *27*(3-4), 109-120.

Mount, M. K., Oh, I. S., & Burns, M. (2008). Incremental validity of perceptual speed and accuracy over general mental ability. *Personnel Psychology*, *61*(1), 113-139.

Appendix I – Short Term (Working) Memory

Definition:

- Short term memory (STM) is a form of temporary storage that does not include mental manipulation of information of input to achieve outputs (level one; Verive & McDaniels, 1996).
  - Note there is a distinction between "level one" and "level two" STM with level two being notably more similar to working memory than level one. Assertions below try to focus on level one.

SGD Reduction Evidence:

- Based on a meta-analysis, Cohen's *d* for various STM tests range from .35 to .58 for Black-White subgroups (favoring Whites; Verive & McDaniels, 1996).
- Verbal STM tests seem to favor women over men (Hough et al., 2001).
- Compared to reading comprehension tests, STM tests seem to produce lower mean score subgroup differences (Barrett et al., 1999).

Incremental Validity Evidence:

- Relationships with performance were strong displaying an observed r of .19 and a corrected r of .41 in a meta-analysis of 31 studies (Verive & McDaniel, 1996). No incremental validity over g was reported, however authors suggest that level 1 STM should be less g-loaded.

References

Barrett, G. V., Carobine, R. G., & Doverspike, D. (1999). The reduction of adverse impact in an employment setting using a short-term memory test. *Journal of Business and Psychology, 14*, 373-377.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*(1-2), 152-194.

Jensen, A. P. (1971). *Do schools cheat minority children?* Paper presented in Seminar Series on Education, Santa Monica, CA.

Verive, J. M. & McDaniel, M. A. (1996). Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence, 23*, 15-32.

Appendix J – Soft Skills "Followership"

Definition:

- Followership is a relatively understudied construct in organizational sciences. Recently it has been re-conceptualized as a role-based construct and as a process-based construct (Uhl-Bien et al., 2014).
  - The role-based approach defines followership as a continuous self-schema ranging from a more passive 'blind obedience' to a more proactive 'change agent' view (Carsten et al., 2010);
  - The process-based approach defines followership in terms of how one becomes a part of the leadership process itself by engaging in the leader-follower relationship through a reciprocal calming and granting of leader and follower identities (Puhl, 2020; Uhl-Bien & Pillai, 2007).
- The military defines followership as Adopts the values and standards of the organization, recognizing one's responsibilities as a follower, and one's role within the organization. Adopts and supports organizational changes. Commits to the action plan of the organization and mission, and advocates for the leader's point of view when a decision is established (Teachout et al., 2019).

SGD Reduction Evidence:

No literature was found

Incremental Validity Evidence:

- No literature was found

References

Carsten, M. K., Uhl-Bien, M., West, B. J., Patera, J. L., & McGregor, R. (2010). Exploring social constructions of followership: A qualitative study. *The Leadership Quarterly*, *21*(3), 543-562.

Puhl, H. D. (2020). *The Skill of Following: Development of a Measure of Followership*. Northern Kentucky University.

Uhl-Bien, M., Riggio, R. E., Lowe, K. B., & Carsten, M. K. (2014). Followership theory: A review and research agenda. *The Leadership Quarterly*, *25*(1), 83-104.

Appendix K – Soft Skills "Growth Mindset"

Definition:

- Growth mindset is the belief that personal characteristics, specifically intellectual ability, are malleable and can be developed by investing time and effort (Dweck, 1999, 2006).

SGD Reduction Evidence:

An unpublished study showed measurement invariance between adolescents and adults (Rammstedt et al., 2021, dissertation).

Incremental Validity Evidence:

- A relatively recent study showed weak negative association between growth mindset and academic achievement (Bahník et al., 2017; Rammstedt et al., 2021, dissertation).
- There is preliminary evidence about a positive association between growth mindset and self-regulation and goal regulation (Rammstedt et al., 2021, dissertation).

References

Bahnik, S., & Vranka, M. A. (2017). Growth mindset is not associated with scholastic aptitude in a large sample of university applicants. *Personality and Individual Differences*, *117*, 139-143.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random house.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development.* Philadelphia: Psychology Press.

Rammstedt, B., Grüning, D. J., & Lechner, C. (2021). *Measuring growth mindset: A validation of a three-item scale and a single-item scale in youth and adults*.

Appendix L – Soft Skills "Instructing"

Definition:

- To be able to teach others how to do something (Mumford et al., 1999; Shore et al., 2019).

SGD Reduction Evidence:

No literature was found

Incremental Validity Evidence:

- No literature was found

References

Mumford, M. D., Peterson, N. G., & Childs, R. A. (1999). Basic and cross-functional skills. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 49–69). American Psychological Association. https://doi.org/10.1037/10313-004

Shore, C. W., Pena, D. A., Gonzalez, M., Haight, N. R., & Wolliston, D. J. (2019). *Officer and Enlisted Needs Analysis* (AFCAPS-FR-2020-0001). JBSA-Randolph AFB, TX: Strategic Research and Assessment Branch, Air Force Personnel Center.

Appendix M – Soft Skills “Moral Courage”

Definition:

- Military defines moral courage as To take action for moral reasons despite the risk of adverse consequences. It is being considered as part of AETC intent "Heart, Mind, & Soul" Ideal Airmen taxonomy (Shore et al., 2019).
- There are five dimensions to moral courage – moral agency, multiple values, endurance of threats, going beyond compliance, and moral goals (Sekerka et al., 2009).

SGD Reduction Evidence:

No literature was found

Incremental Validity Evidence:

- No literature was found

References

Sekerka, L. E., Bagozzi, R. P., & Charnigo, R. (2009). Facing ethical challenges in the workplace: Conceptualizing and measuring professional moral courage. *Journal of Business Ethics*, *89*(4), 565-579.

Shore, C. W., Pena, D. A., Gonzalez, M., Haight, N. R., & Wolliston, D. J. (2019). *Officer and Enlisted Needs Analysis* (AFCAPS-FR-2020-0001). JBSA-Randolph AFB, TX: Strategic Research and Assessment Branch, Air Force Personnel Center.

List of Symbols, Abbreviations, and Acronyms

%	percent
a parameter	Item Response Theory Discriminability Parameter
A/W	Asian/White Comparison
AA1 & 2	Academic Assessment 1 & 2
ABM	Air Battle Manager
ABMMASSTotal	ABM Merit Assignment Selection System Total for Primary Training
AFOQT	Air Force Officer Qualifying Test
AFPC/DSYX	Air Force Personnel Center Strategic Research and Assessments Branch
AI	Aviation Information
ANIT	Aviation and Nautical Information test
AO	Assembling Objects
AR	Arithmetic Reasoning
ASTB	Aviation Selection Test Battery
ASVAB	Armed Services Vocational Aptitude Battery
Avg	Average
b parameter	Item Response Theory Difficulty Parameter
B/W	Black/White Comparison
B1	Informative Briefing
B2	Advocacy Briefing
BC	Block Counting
CD	Compass Directions Test
Cohen's *d*	Effect Size
CombMASSTotal	Combined Merit Assignment Selection System Total for Primary Training
CombRawAcad	Combined Raw Academic Score for Primary Training
CombRawCheck	Combined Raw Flight Check Score for Primary Training
CombRawDaily	Combined Raw Daily Score for Primary Training
CombRawFltCC	Pilots Raw Flight Check
α	Cronbach's Alpha
CSO	Combat Systems Officer
CSOMASSTotal	CSO Merit Assignment Selection System Total for Primary Training
CTT	Classical Test Theory
DoD	Department of Defense
DRT	Dial Reading Test
ECAT	Enhanced Computer Administered Test
EDPT	Electronic Data Processing Test
F/M	Female/Male Comparison
FLTCHK	Flight Checking Test
FW	Figure Weights
g	General Menal Ability
GLP 1 & 2	Graded Leader Position 1 and 2
GPA	Grade Point Average
GS	General Science
IC	Instrument Comprehension
IFT	Initial Flight Training

IRT	Item Response Theory
ITC	Item-Total Correlation
Kurt.	Kurtosis
LRC	Leader Reaction Course
M	Mean
MASS	Merit Assignment Selection System
Max	Maximum
MC	Mechanical Comprehension
MCt	Mental Counters
MEML	Memory for Landmarks
Min	Minimum
MK	Math Knowledge
MST	Math Skills Test
N/A	Not Applicable
NECFACT	Necessary Facts Test
NS	Number Series
NVR	Non-Verbal Reasoning
OSM	Officer Suitability Measure
OTS	Officer Training School
P1	Informative Paper
P2	Advocacy Paper
PBM	Performance Based Measure
PS	Physical Science
p-value	Proportion of correct items compared to all items in the assessment
RC	Reading Comprehension
RCT	Reading Comprehension Test
RIQ	Instrument Qualification
SD	Standard Deviation
SDI-O	Self-Description Inventory - Officers
SGDs	Subgroup Differences
SJT	Situational Judgment Test
Skew.	Skewness
SUPer	Selection of Unmanned Aerial Systems Personnel
SUPT	Specialized Undergraduate Pilot Training
TR	Table Reading
TBAS	Test of Basic Aviation Skills
UAS	Unmanned Aerial Systems
US	United States
USAF	United States Air Force
VA	Verbal Analogies
WELPS	Warrior Expeditionary Leadership Problem Solving
WK	Word Knowledge
WnH/WH	White Non-Hispanic/White Hispanic comparison
WPt	Weight perception

[i] Cohen's *d* = standardized mean score difference.