AFRL-RH-WP-TR-2024-0003

AIR FORCE OFFICER QUALIFYING TEST FORM U: REVIEWING COGNITIVE SUBTESTS FROM PREVIOUS FORMS

Julia L. Walsh, Rusty Wilson, Kyle J. Mann
Infoscitex, a DCS Corp

John Trent
Air Force Personnel Center
Strategic Research and Assessment Branch

Thomas R. Carretta
711 HPW/RHBC

February 2024
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
HUMAN EFFECTIVENESS 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-2024-0003 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: 2024.03.04 15:27:48 -05'00'

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: 2024.03.12 13:13:52 -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*)**
08-01-24	Interim	September 2021 – January 2022

4. TITLE AND SUBTITLE

Air Force Officer Qualifying Test Form U: Reviewing Cognitive Subtests from Previous Forms

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], John Trent[b], and Thomas R. Carretta[c]

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H12Q

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
[a]Infoscitex, a DCS Corporation
4027 Colonel Glenn Highway, Suite 210
Dayton, OH 45431-1672

8. PERFORMING ORGANIZATION REPORT NUMBER

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

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

[b]Air Force Personnel Center
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-2024-0003

12. DISTRIBUTION AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Report contains color. AFRL-2024-1217, cleared 4 March 2024

14. ABSTRACT

This report describes empirical, theoretical, and competency evaluation of the 17 cognitive subtests included in the Air Force Officer Qualifying Test (AFOQT) Forms O, P, Q, R/S, and T. Empirical evaluation involved a review of archival information, such as technical reports and test manuals, to document the psychometric properties of each subtest to determine whether or not to include it in the AFOQT Form U. Theoretical evaluation involved a review of the academic literature on the dominant cognitive ability frameworks, including the Cattell-Horn-Carroll (CHC) model, to ensure alignment between the AFOQT and the theoretical frameworks. Finally, competency evaluation involved a mapping between the subtests and the critical competencies that the United States Air Force (USAF) officers need to have upon commissioning and upon classifying into rated (i.e., aircrew) career fields. Taken together, the findings led to the following recommendations: (1) do not bring back any de-commissioned subtests; (2) exclude Word Knowledge due to its poor psychometric performance and poor linkage to the USAF officer competencies; and (3) modify some operational AFOQT Form T subtests as described in this report. Because the mapping between some of the operational subtests and the core officer competencies was tenuous at best, we recommend that the list of the subtests be augmented with some measures of fluid intelligence or specific abilities. We are confident that these enhancements would improve the psychometric integrity of the AFOQT as a whole.

15. KEY WORDS
Air Force Officer Qualifying Test, AFOQT, empirical evaluation, theoretical evaluation, competency mapping

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Thomas R. Carretta
Unclassified	Unclassified	Unclassified	SAR	39	**19b. TELEPHONE NUMBER** *(Include Area Code)*

TABLE OF CONTENTS

LIST OF TABLES

1.0 EXECUTIVE SUMMARY

The Air Force Personnel Center Strategic Research and Assessments Branch (AFPC/DSYX) initiated a contract to evaluate the 17 cognitive subtests included in the previous Forms (O, P, Q, R/S, and T) of the Air Force Officer Qualifying Test (AFOQT) to determine which ones should be considered for 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 tradeoff; Ployhart & Holtz, 2008). Note that of the 17 cognitive subtests, 10 are in the AFOQT Form T and seven are no longer operational.

A three-pronged technical approach was adopted to conduct an empirical, theoretical, and competency evaluation of the 17 cognitive subtests included in the AFOQT Forms O, P, Q, R/S, and T. The empirical evaluation involved a review of the archival information, such as technical reports and test manuals, to document the psychometric properties of these subtests. The review relied on general subtest information, meta-data, Classical Test Theory (CTT) data, and criterion-related data. The theoretical evaluation involved a review of the academic literature on the dominant cognitive ability frameworks (i.e., models of intelligence), namely Cattell's Crystallized and Fluid Intelligence (C&F) model, Cattell-Horn-Carroll (CHC) model, and Thurstone's Primary Mental Abilities (PMA) model, to ensure the alignment between the cognitive subtests and the theoretical frameworks. Finally, the competency evaluation involved a review of the ongoing and completed efforts on the critical competencies that USAF officers need to have upon commissioning and classifying into rated (i.e., aircrew) career fields, such as manned and remotely piloted aircraft (RPA) pilot, combat systems officer (CSO), and air battle manager (ABM). This step was necessary to determine how well the subtests cover key USAF officer competencies and whether they add predictive power above and beyond the other cognitive subtests.

From an empirical perspective, there are several takeaways. First, there is evidence not to bring back any de-commissioned subtests. While some of them exhibited acceptable or even favorable criterion-related validities, almost all of the CTT metrics indicated deficiencies (most notably the mean score subgroup difference [SGD] values). Second, most operational subtests are in an acceptable standing, with CTT metrics being mildly concerning or favorable. This suggests that these subtests can be included in the AFOQT Form U but will need to be modified slightly, focusing specifically on the concerning metrics (e.g., difficulty, discriminability, SGD). Third, the rest of the operational subtests can be included in Form U but with substantial modifications. Finally, Word Knowledge (WK) should be excluded from the next generation AFOQT battery for its poor CTT metrics and criterion-related validities.

From a theoretical perspective, the main takeaway is that the 17 cognitive subtests measure crystallized intelligence to a great extent, fluid intelligence to some extent, and other types of intelligence to a small extent or not at all. While not all types of intelligence are applicable to the goals of the AFOQT, there were a few that could have been measured better (e.g., short-term memory, long-term storage and retrieval, and reaction and decision speed). With that said, note that a greater emphasis was given to the alignment between the cognitive subtests and the officer competencies than to the alignment between the cognitive subtests and the theoretical frameworks.

From a competency evaluation perspective, the results were mixed. On the one hand, (1) only two de-commissioned subtests mapped strongly to the core officer competencies while the rest of the operational and de-commissioned subtests mapped to the competences only moderately or not at all; and (2) most of the operational and de-commissioned subtests mapped to the Air Force Specialty Code (AFSC)-specific competencies only moderately or not at all. WK was the only operational subtest that mapped only moderately to the aircrew career fields competencies, while the rest of the operational subtests mapped strongly. Furthermore, three operational subtests did not survive the stepwise regression model indicating a lack of incremental predictive power.

Considering the results of the aforementioned evaluations, we recommend that the next generation AFOQT Form U reimagines its suite of the cognitive ability subtests by (1) not bringing back any de-commissioned subtests; (2) excluding WK due to its poor psychometric performance and poor linkage to the USAF officer competencies; and (3) modifying the rest of the operational subtests as described in this report. Because the mapping between some of the operational subtests and the core officer competencies was tenuous at best, we recommend that the subtests be augmented with measures of fluid intelligence and/or specific abilities. We are confident that these enhancements would improve the psychometric integrity of the AFOQT as a whole.

2.0 INTRODUCTION

The effort described in this technical report served as the first step in preparing the AFOQT Form U. The main goal of this effort was to evaluate the 17 cognitive subtests included in the AFOQT Forms O, P, Q, R/S, and T to determine which of them should be included in the Form U. The next steps in preparing Form U are described in other technical reports. The following paragraphs provide a brief history of the AFOQT and review the evolution of the Forms and subtests throughout the years.

2.1 Brief History of the AFOQT

The AFOQT has been an important component of the Air Force Personnel Testing Program (AFPTP) since 1953. It is a critical tool for officer commissioning and aircrew training classification and is widely accepted among military personnel selection communities as a useful and cost-effective instrument. Historically, the AFOQT has been the primary selection test for the Air Force Reserve Officer Training Corps (AFROTC), Officer Training School (OTS), and the Airman Education and Commissioning Program (AECP). It is also used in the selection process for Undergraduate Pilot Training (UPT), Undergraduate RPA Training (URT), CSO training, and ABM training. Since its inception, the AFOQT has undergone several revisions to improve its performance prediction and officer classification (for a complete history of the AFOQT see Drasgow et al., 2010).

2.2 AFOQT Forms O, P, Q, R/S, and T

Although the AFOQT has been operational since 1953, the current effort focused on the 17 cognitive subtests included in Forms O, P, Q, R/S, and T only (see Table 1). As can be seen, Form R was never administered, so the administration went from Form Q straight to Form S. Form S incorporated a non-cognitive subtest known as the Self-Description Inventory Plus (SDI+), which was designed to measure officer personality traits. Subsequently, a modified version of the SDI+, known as SDI-O (Self-Description Inventory - Officers), was incorporated into Form T along with an experimental situational judgment test (SJT). The evaluation of the non-cognitive subtests is beyond the scope of the current report. For more information on the SDI+ and SDI-O, please refer to Mann et al. (2023) and Woolley et al. (2022). For more information on the SJT, please refer to Walsh et al. (2022) and Sizemore et al. (2022).

Table 1. Overview of the AFOQT Forms

Form	Years of Administration	Number of Parallel Forms	Number of Cognitive Subtests	Administration Time*
O	1981 - 1987	1	16	3.5 hours
P	1987 - 1994	2	16	3.5 hours
Q	1994 - 2005	2	16	3.5 hours
R	[not administered]	2	16	N/A
S	2005 - 2015	2	11	2 hours
T**	2015 - present	2	10	2.3 hours

*The administration time column does not include the time allotted for non-cognitive subtests (e.g., Situational Judgment Test, Self-Description Inventory Plus, Self-Description Inventory for Officers).
**Form T Version 1 was deployed in 2015. Form T Version 2 was deployed in 2023 (Kantrowitz et al., 2023).
Note. N/A = Not Applicable.

The 17 cognitive subtests evaluated in this effort are summarized in Table 2. For item samples please refer to Attachment 1.

Table 2. Description of the AFOQT Cognitive Subtests

Operational Subtests		
Subtest	Abbreviation	Description
Verbal Analogies	VA	Measures the ability to reason and recognize relationships between words. The examinee must choose the option that best completes the analogy developed at the beginning of each statement.
Arithmetic Reasoning	AR	Measures the understanding of arithmetic relationships expressed as word problems. Each problem is followed by five possible answers. The examinee must decide which one of five choices is correct.
Word Knowledge	WK	Measures the ability to understand written language through the use of synonyms. For each question, the examinee must choose the word that is closest in meaning to the capitalized word provided.
Math Knowledge	MK	Measures knowledge of mathematical terms, formulas, and relationships. Each problem is followed by five possible answers. The examinee must decide which one of the five choices is correct.
Reading Comprehension	RC	Measures the ability to read and understand written material. Each passage is followed by a series of multiple-choice questions. The examinee must choose the option that best answers the question based on the passage. No additional information or specific knowledge is needed.
Physical Science	PS	Measures knowledge in the area of science. Each of the questions or incomplete statements is followed by five

		choices. The examinee must decide which of the choices best answers the question or completes the statement.
Table Reading	TR	Measures the ability to read a table quickly and accurately. For each test question, the examinee is given an X-axis and a Y-axis value. The examinee's task is to find the cell where the column and row intersect, note the number that appears there, and then find this number among the five provided answer options.
Instrument Comprehension	IC	Measures the ability to determine the position of an airplane in flight from reading instruments showing its compass direction heading, amount of climb or dive, and degree of bank to right or left. Each problem consists of two dials and four airplanes in flight as answer options. The examinee's task is to determine which one of the four airplanes is most nearly in the position indicated by the two dials.
Block Counting	BC	Measures the ability to "see into" a 3-dimensional pile of blocks. Given a certain numbered block, the examinee's task is to determine how many other blocks the numbered block touches.
Aviation Information	AI	Measures knowledge of aviation. Each of the questions or incomplete statements is followed by five choices. The candidate is to decide which one of the choices best answers the question or completes the statement.
De-Commissioned Subtests		
Subtest	Abbreviation	Description
Data Interpretation	DI	Measures the ability to extract data from graphs and charts
Mechanical Comprehension	MC	Measures the understanding of mechanical functions
Electrical Maze	EM	Measures spatial ability based on choice of a path through a maze
Rotated Blocks	RB	Measures spatial ability by requiring mental manipulation and rotation of objects
Hidden Figures	HF	Measures spatial ability by requiring the discovery of simple figures embedded in a complex figure
General Science	GS	Measures knowledge and understanding of scientific concepts, terms, principles, and instruments
Scale Reading	SR	Measures the ability to read scales and dials.

Table 3 summarizes the status of the 17 AFOQT cognitive subtests throughout the years. As can be seen, some subtests remained the same, some changed slightly, and others were dropped/de-commissioned altogether. Eight subtests have been consistently included in AFOQT Forms O through T: VA, AR, WK, MK, IC, BC, TR, and AI. The content taxonomy for GS was changed with the release of Form T, when it was renamed PS and focused directly on physical sciences (e.g., chemistry and physics; Carretta et al., 2016). RC, DI, MC, EM, and SR were removed when

Form S was implemented due to their lack of incremental validity (Parish et al., 2008) and the desire to broaden AFOQT content with a measure of personality (SDI+). RB and HF were removed from AFOQT Form T due to their lack of incremental validity (Johnson et al., 2017). Note that RC was removed from Form S due to the time concerns and the fact that it did not add incremental validity beyond other verbally loaded subtests (VA and WK; Parish et al. 2008). However, after Form S was implemented, there was a concern about the examinees' reading level, so RC was brought back.

Table 3. Status of the AFOQT Cognitive Subtests between Forms

Form O	Form P	Form Q	Form R	Form S	Form T
VA	VA	VA	VA	VA	VA
AR	AR	AR	AR	AR	AR
WK	WK	WK	WK	WK	WK
MK	MK	MK	MK	MK	MK
IC	IC	IC	IC	IC	IC
BC	BC	BC	BC	BC	BC
TR	TR	TR	TR	TR	TR
AI	AI	AI	AI	AI	AI
GS	GS	GS	GS	GS	PS
RC	RC	RC	RC	*dropped*	RC
RB	RB	RB	RB	RB	*dropped*
HF	HF	HF	HF	HF	*dropped*
MC	MC	MC	MC	*dropped*	*dropped*
EM	EM	EM	EM	*dropped*	*dropped*
SR	SR	SR	SR	*dropped*	*dropped*
DI	DI	DI	DI	*dropped*	*dropped*

The inclusion and exclusion of the various subtests over the years was accompanied by subtest-level modifications in terms of the number of items, the number of response options, and the administration times (see Table 4). Note that of the subtests included in Form T, IC, PS, BC, and RC underwent the greatest number of changes.

Table 4. Changes to the AFOQT Cognitive Subtests

Subtest	Number of Items	Number of Response Options	Administration Time
VA	25	5	8 min
AR	25	5	29 min
WK	25	5	5 min
MK	25	5	22 min
IC	20 to 25	4 to 5	6 to 5 min
BC	20 to 30	5	3 to 4.5 min
TR	40	5	7 min
AI	20	5	8 min
GS/PS	20	5	10 min
RC	25	5	18 to 38 min
RB	15	5	13 min
HF	15	5	8 min
MC	20	5	22 min
EM	20	5	10 min
SR	40	5	15 min
DI	25	4 to 5	24 min

3.0 METHOD

The previous section focused on reviewing the evolution of the AFOQT battery throughout the years to understand how and why the 17 cognitive subtests were changing. The following paragraphs describe the technical approach that was undertaken to evaluate the cognitive subtests to determine which of them should be included in AFOQT Form U.

3.1 Technical Approach

A three-pronged technical approach was adopted: (1) an empirical evaluation, (2) a theoretical evaluation, and (3) a competency evaluation.

3.1.1. Empirical Evaluation

We searched Defense Technical Information Center (DTIC) and Google Scholar for technical reports, test manuals, and journal articles about the AFOQT cognitive subtests. We also gathered some of the materials from the AFPC/DSYX. Finally, we used internal knowledge and the technical reports prepared by Infoscitex (IST) and Personnel Decisions Research Institute (PDRI) regarding Form T (Kantrowitz et al., 2022; Walsh et al., 2022; Walsh et al., 2021). Data gathered can be grouped into five categories: (1) general information, (2) meta-data, (3) CTT data, (4) criterion-related validity data, and (5) Item Response Theory (IRT) data.

General information focused on documenting the following information for each subtest: (1) to which AFOQT Form(s) it belonged, (2) a brief description of its content, (3) the constructs it purported to measure, (4) whether it was knowledge-based or aptitude-based, (5) whether it was used to qualify for one or more rated career fields, and (6) the years of administration.

Meta-data focused on which composites each subtest contributed to (Pilot, ABM, CSO, Verbal, Quantitative, and/or Academic Aptitude), how many items and response options it contained, and how many common items it had between the parallel forms.

CTT data focused on the available test statistics: (1) descriptives (i.e., mean [*M*], standard deviation [*SD*], skewness); (2) difficulty (i.e., *p*-value); (3) speededness (yes or no); (4) discriminability (i.e., item-total correlations [ITC]); (5) internal consistency (i.e., Cronbach's alpha); and (6) mean score subgroup differences (i.e., effect sizes expressed as Cohen's *d*). Note that the main subgroups of interest to this research were gender, race, and ethnicity. Also note that not all reports included this information, therefore we made comparisons between and within Forms based on the available data. See Appendix A for more information about the CTT metrics.

Criterion-related data focused on the available validities between the AFOQT cognitive subtests and various USAF outcomes (e.g., Airmanship, AFROTC Grade Point Average [GPA]). Note that the criteria were not consistent between Forms making it challenging to make fair comparisons. Thus, evaluations of criterion-related validity were made with caution.

Finally, IRT data focused on available item-level 2-parameter logistic (2PL) and 3-parameter logistic (3PL) parameter estimates generated for the relevant subtests (i.e., speeded tests are not amenable to unidimensional IRT analyses). Note that IRT metrics were available only for Forms O and T. See Appendix A for more information about the IRT metrics.

When deciding which subtests to recommend for inclusion in Form U, our team of industrial-organizational (I/O) psychologists evaluated the subtests wholistically both between and within Forms. To make evaluations, we devised a set of cutoffs against which the metrics were compared (see Appendix B). Our expectation was that the operational subtests would exhibit stronger psychometric properties compared to the de-commissioned subtests. However, we also expected to find opportunities to improve the operational subtests.

3.1.2. Theoretical Evaluation

It is well documented in the literature that general mental ability (*g*) predicts important organizational outcomes, such as task performance, training performance, organizational citizenship behaviors, and counterproductive work behaviors (Schmidt & Hunter, 2004; Schneider & Newman, 2015). The general factor is usually heralded as the "king" of personnel selection assessments, so much so that some scholars argue that there is not much more than *g* in terms of predictive validity (Jensen, 1998; Ree & Carretta, 2022).

However, *g*-loaded tests consistently result in moderate to large mean score SGDs for gender, racial, and ethnic subgroups (a phenomenon known as the validity-diversity tradeoff; Ployhart & Holtz, 2008). With many organizations, including the USAF, emphasizing diversity, research has

shifted toward exploring the power of specific abilities (*s*). The growing body of research suggests that *s* increments predictions above and beyond *g* by approximately 2% (Lang et al., 2010; Ree & Carretta, 2022). Although this value appears low, it may have wide-ranging benefits in terms of practical utility and increased diversity. The bandwidth-fidelity dilemma suggests that '*s'* may be more appropriate for predicting specific performance outcomes, whereas '*g'* is more appropriate for predicting general performance outcomes (e.g., Ones & Viswesvaran, 1996). Additionally, positive manifold suggests that specific abilities themselves are important beyond their predictive ability due to the tendency for specific abilities to build upon one another and develop in tandem (van der Maas et al., 2006). Together the theoretical and empirical evidence suggests that it is important to develop assessments that contain both broad and narrow types of intelligence. Therefore, we examined the prevailing theories of intelligence to determine if there is contamination or deficiencies in the AFOQT assessment.

Our literature review identified three prominent models of intelligence. The first is the model of C&F intelligence devised by Cattell (1941, 1943, 1950). This model denotes a distinction between crystallized intelligence (Gc) (i.e., a reliance on prior knowledge or experience to solve problems; McDaniels & Banks, 2010) and fluid intelligence (Gf) (i.e., an ability to solve novel problems through reasoning; McDaniels & Banks, 2010). The second is the CHC model of intelligence that suggests intelligence is multifaceted and hierarchical, being comprised of four higher-order factors (knowledge, controlled attention, perception, and motor) which can be divided into 16 lower-order factors[1]. Finally, Thurstone's PMA model outlines seven factors of intelligence (word fluency, verbal comprehension, numerical ability, visual-spatial, perceptual speed, memory, and inductive reasoning). See Appendix C for more information on the three theoretical frameworks used in this effort.

Three I/O psychologists mapped each cognitive subtest to the constructs within each of the three intelligence models. After doing that, we evaluated the findings wholistically. Our expectation was that the subtests would map to the CHC model better than to the other two models because the AFOQT (as do many personnel selection and classification assessments) typically follow the CHC model. However, we also expected to find opportunities for improvement.

3.1.3. Competency Evaluation

In addition to the empirical and theoretical evaluation, it was important to evaluate how well the 17 cognitive subtests cover the key attributes and competencies necessary for successful performance in USAF jobs. To accomplish this goal, we reviewed the consolidated competency model developed by Persing et al. (in press), results of a mapping task for the AFOQT Form T and the new composite calculations for the AFOQT Form T (Kantrowitz et al., 2022).

Persing et al. (in press) reviewed and consolidated the various USAF needs analyses and competency models published in recent years. The consolidated model contained 15 core competencies and 29 occupation-specific (i.e., AFSC) competencies. The core competencies are

[1] The authors are aware of the research comparing the CHC model against Carroll's Three Stratum Theory and that the latter model is deemed more empirically robust than the former model (Cucina & Howardson, 2017). However, there is significant overlap between the two models, which would yield similar mapping results. Because the CHC model is widely accepted, we decided to use it instead of Carroll's model.

applicable to all ranks of enlisted and officers. The AFSC-specific competencies are uniquely applicable to some occupations. For the current effort, we asked four I/O psychologists to link the 44 competencies to the 17 AFOQT cognitive subtests. The I/O psychologists were instructed to rate the extent to which a cognitive subtest measured a given competency using a 0-2 Likert-type scale, where 0 was 'not at all,' 1 was 'related to a moderate extent,' and 2 was 'related to a great extent.' Average ratings of below 1 were considered weak, between 1 and 1.49 were considered moderate, and above 1.50 were considered strong.

A separate effort by Kantrowitz et al. (2022) used needs analyses for rated career fields only (CSO, mobility pilots, fighter pilots, and RPA pilots) to link 67 competencies and attributes to the 10 cognitive and two non-cognitive AFOQT Form T subtests. Note that the information on ABMs was not available. The researchers used the same scale as described above. As a result of this effort, each cognitive subtest received a score between 0 and 2 indicating how well it measured the attributes and competencies for the aforementioned rated career fields. Measurement gaps were identified.

Kantrowitz et al. (2022) performed Pareto-Optimization and regression-based analyses for the 10 operational cognitive subtests and two experimental non-cognitive subtests on the AFOQT Form T to determine whether each subtest should be included in the current and/or alternative officership and aircrew aptitude composites based on its ability to optimize the diversity-validity tradeoff while maintaining predictive validity. As a result of this effort, some subtests were flagged for inclusion while others were flagged for exclusion.

When deciding which subtests to recommend for inclusion in Form U, we evaluated the three aforementioned efforts wholistically. Our expectation was that operational subtests would link to the core and AFSC-specific competencies better than the de-commissioned subtests. We also expected that the operational subtests would be included in the AFOQT composites. We also expected to find opportunities for improvement.

4.0 RESULTS

The previous section overviewed the technical approach undertaken to evaluate the 17 cognitive subtests included in the AFOQT Forms O, P, Q, R/S, and T to determine which of them should be included in the AFOQT Form U. The following paragraphs detail the results of the empirical, theoretical, and competency evaluations.

4.1 Empirical Evaluation

Table 5 contains the cognitive subtests and their evaluation across the CTT metrics and criterion-related validities. Color codes indicate the following: green – information was favorable, orange – information was acceptable, and red – information was concerning. See Attachment 2 for a complete summary of the psychometrics reviewed in this effort. Recall that we expected that the operational subtests would exhibit stronger psychometric properties compared to the de-commissioned subtests. However, we also expected to find opportunities for improvement.

Table 5. Results of Empirical Evaluation

Subtest	Difficulty	Discriminability	Internal Consistency	F/M SGD	B/W SGD	A/W SGD	H/nH SGD	Criterion Validity
VA					**	**	**	**
AR								*
WK				**	**	**	**	**
MK				**	**		**	**
RC		*		**	**	**	**	**
PS	**	**	**	**	**	**	**	**
TR		*			**	**	**	**
IC				**	**	**	**	**
BC								*
AI								*
DI		*		*	*	*	*	*
MC		*		*	*	**	**	*
EM		*		*	*	*	*	*
RB		*		*	*	*	*	*
HF		*		*	*	*	*	*
GS				*	*	*	*	*
SR		*	*	*	*	*	*	*

Note. F/M SGD = Female/Male mean score SGDs, B/W SGD = Black/White mean score SGDs, A/W SGD = Asian/White mean score SGDs, H/nH SGD = Hispanic/Non-Hispanic mean score SGDs. The underlined subtests appear in the AFOQT Form T.
*Data were available for only one or two AFOQT Forms.
**Data were available only for AFOQT Form T.

First, considering that the AFOQT was designed to target average-to-below-average ranges of cognitive ability, most subtests performed well; two subtests were too difficult (AI and EM) and five subtests were too easy (VA, RC, TR, DI, and HF). The reason for coding AI and EM in green (and not in red) is because between easy subtests and difficult subtests, the latter may be more preferred because the former may result in ceiling effects and therefore may waste valuable resources without any gains in information about the candidates. In the case of the AFOQT, average to below-average range of difficulty should be targeted (closer to the orange range).

Second, most subtests had average discriminability, suggesting that they can discriminate between candidates' ability levels (low to average to high) fairly well. Five subtests (WK, IC, DI, EM, and HF) had great discriminability. One subtest (VA) had poor discriminability.

Third, internal consistency was adequate or favorable for all subtests.

Fourth, most of the operational AFOQT Form T subtests had lower mean score SGDs compared to those included in previous Forms. However, there were still large mean score SGDs for the Black/White (B/W) subgroups and moderate mean score SGDs for the gender and ethnic

subgroups. An important caveat to this observation is that the literature that we reviewed had little information about the mean score SGDs across the AFOQT Forms, precluding us from deriving a more balanced review of the metrics over time. This is explained by the fact that while diversity and inclusion were important throughout the years, it was not until recently that the reporting of the effect sizes became routine.

Finally, most of the subtests exhibited favorable or acceptable criterion-related validities. WK and MK were the only two subtests that showed weak validities throughout the years.

Overall, our expectations were met in that operational subtests showed stronger psychometric properties compared to the de-commissioned subtests. Furthermore, we found evidence not to bring back any de-commissioned subtests. While some of them exhibited acceptable or even favorable criterion-related validities, almost all of the other CTT metrics were mildly or very concerning (most notably the mean score SGD effect sizes). Second, most operational subtests were in an acceptable standing, with most metrics being mildly concerning or favorable. This suggests that these subtests, namely AR, RC, TR, and IC, can be included in the AFOQT Form U but will need to be modified slightly, focusing specifically on difficulty, discriminability, and SGD. Third, the rest of the operational subtests, namely VA, MK, PS, BC, and AI, can be included in Form U but with substantial modifications. Finally, WK should be excluded from the next generation AFOQT battery for its poor CTT metrics and criterion-related validities.

4.2 Theoretical Evaluation

Table 6 summarizes the evaluation of the 17 cognitive subtests against the three prevailing theories of intelligence to determine whether there was contamination and/or deficiency in coverage. The color green indicates that the subtests cleanly mapped onto a single type of intelligence and the type of intelligence is a sufficient fit for the subtest. Orange indicates that the subtest mapped onto multiple types of intelligence or only partially mapped onto a single type of intelligence. Red indicates that the subtest did not map onto any type of intelligence. Recall that we expected that the subtests would map to the CHC model better than to the other two models. However, we also expected to find opportunities for improvement.

When examining the C&F model, most subtests cleanly mapped onto either Gc or Gf. However, six subtests (i.e., VA, RC, IC, MC, DI, SR) mapped onto both types of intelligence, which may be due to the overlap in the cognitive models. Note that three of these subtests were included in the AFOQT Form T (i.e., VA, RC, and IC). Also note that there were more Form T operational subtests that mapped to Gc than to Gf, which arguably can explain nontrivial mean score SGDs.

When examining the CHC model, we mapped the subtests only to the types of intelligence that were applicable to the goals of the AFOQT. The types of intelligence that were not applicable included auditory processing, olfactory processing, tactile processing, kinesthetic processing, psychomotor abilities, and psychomotor speed. This left 10 types of intelligence to map. Of these, seven mapped to the subtests, with some types mapping only to one subtest and others cross-mapping to more than one subtest, suggesting an overlap between the models. Three types of intelligence that were not mapped to any subtests included short term memory, long-term storage and retrieval, and reaction and decision speed. This suggests potential measurement deficiency.

Note that the CHC and the C&F models converged on the finding that more Form T operational subtests mapped to Gc (domain-specific knowledge) than to Gf.
Regarding the PMA model, most of the subtests covered the intelligence types well, however, three subtests did not map to any intelligence types. This suggests either contamination or a misalignment between the AFOQT and the PMA model, which is not surprising given that most of the personnel selection and classification assessments tend to follow the CHC model.

Table 6. Results of Theoretical Evaluation

Subtest	C&F Model	CHC Model	PMA Model
VA			
AR			
WK			
MK			
RC			
PS			
TR			
IC			
BC			
AI			
DI			
MC			
EM			
RB			
HF			
GS			
SR			

Note. Underlined subtests appear in Form T. C&F = Crystallized and Fluid Intelligence; CHC = Cattell-Horn Carroll; PMA = Thurstone's Primary Mental Abilities.

Overall, our expectations were met in that most subtests mapped better to the CHC model than the other two models. Across the three models however, a common observation was that the cognitive subtests measured Gc to a great extent, measured Gf to some extent, and measured other types of intelligence to a small extent or not at all. As a result, we recommend that the next generation AFOQT be more closely aligned with the identified USAF officer attributes and competencies first and to the theoretical models second. Appendix D presents the list of the missing or not well-measured types of intelligence and provides ways to measure them.

4.3 Competency Evaluation

Table 7 displays the subtests evaluated across the previous efforts (Kantrowitz et al., 2022; Persing et al., in press). Green indicates subtests that strongly linked to competencies and attributes or that were included in the composites (based on the results of the regression and Pareto-Optimization analyses). Orange indicates subtests that moderately linked to competencies and attributes. Red

indicates subtests that linked either poorly or not at all to competencies and attributes and were not included in the composites (based on the results of the regression and Pareto-Optimization analyses). Appendix E presents the results of the mappings. Recall that we expected that the operational subtests would link to the core and AFSC-specific competencies better than the de-commissioned subtests. We also expected that the operational subtests would be included in the AFOQT composites. As previously discussed, we also expected to find opportunities for improvement.

Table 7. Results of Competency Evaluation

Subtest	Persing et al. (in press) Core Officer Competencies	Persing et al. (in press) AFSC-Specific Officer Competencies	Kantrowitz et al. (2022) Rated Career Field Competencies	Kantrowitz et al. (2022) Composites
VA				
AR				
WK				
MK				
RC				
PS				
TR				
IC				
BC				
AI				
DI			N/A	N/A
MC			N/A	N/A
EM			N/A	N/A
RB			N/A	N/A
HF			N/A	N/A
GS			N/A	N/A
SR			N/A	N/A

Note. Underlined subtests appear in Form T. N/A = Not Applicable; AFSC = Air Force Specialty Code.

In terms of core competencies, among operational subtests, none had strong linkages to these competencies, seven had moderate linkages to at least one competency, and three (i.e., WK, AI, PS) had no linkages whatsoever. Among de-commissioned subtests, two (i.e., RB, DI) linked strongly to Problem Solving, four (i.e., EM, SR, RB, DI) had moderate linkages, and three (i.e., GS, HF, MC) had no linkages at all. In terms of AFSC-specific competencies, among operational subtests, none had strong linkages to these competencies, four had moderate linkages (i.e., VA, WK, IC, TR), and six had no linkages at all. Among de-commissioned subtests, two had moderate linkages (i.e., EM, SR) and the rest had no linkages. The results are surprising in that (1) the general aptitude operational subtests did not link strongly to the core competencies which are necessary for all USAF officer personnel and that (2) the more specialized subtests (e.g., AI, PS) did not link

strongly or at all to the AFSC-specific competencies. This indicates potential deficiency in measurement which could lead to poor predictive validity. Appendix F presents the list of the core and AFSC-specific competencies that are currently unmeasured and provides ways to measure them.

Kantrowitz et al. (2022) reported that nine out of 10 AFOQT Form T cognitive subtests linked strongly to the attributes and competencies necessary to be successful in the aircrew career fields. WK received only moderate linkage. These findings are encouraging considering that both general aptitude and specialized subtests make up operational composites that are used in criterion-related validation studies. An important observation is that PS (an experimental subtest) was strongly linked to some competencies indicating its utility for prediction in the future. Kantrowitz et al. provided a list of the competencies that were unmeasured or were not measured well by the operational subtests and offered examples of ways to measure them.

When examining the existing operational composites and experimental composites, three subtests (VA, WK, and BC) did not survive the stepwise regression process and Pareto Optimization process. This places these subtests at risk because they do not increment prediction above and beyond the rest of the cognitive subtests and do not help minimize mean score SGDs.

Overall, the results from these efforts were mixed. On the one hand, the operational subtests either did not map strongly or at all to the core and AFSC-specific competencies painting a fairly grim projection for the criterion-related validation studies. On the other hand, nine of 10 subtests mapped strongly to the rated career field competencies painting a bright projection for the criterion-related validation studies. The most perplexing is the finding that VA, WK, and BC did not contribute to the composites suggesting that they may add little value in the future. With that said, the one clear observation is that WK does not seem to map to the competencies across independent efforts and does not contribute to the composites.

5.0 DISCUSSION AND RECOMMENDATIONS

A project was initiated to evaluate the 17 cognitive subtests included in the AFOQT Forms O, P, Q, R/S, and T to determine which should be included in AFOQT Form U. This project is part of a continuous effort by the USAF to predict important personnel outcomes while increasing qualification rates for historically underrepresented gender, racial, and ethnic minority subgroups.

A three-pronged technical approach was adopted to conduct an empirical, theoretical, and competency evaluation of the cognitive subtests. From an empirical perspective, there are several takeaways. First, there is evidence not to include any de-commissioned subtests in future AFOQT Forms. While some of these subtests exhibited high criterion-related validity, almost all other metrics were mildly or very concerning (most notably the mean score SGD values). Substantial time and effort would be needed to improve these subtests which may not be the best use of resources. Second, although most of the operational subtests were in an acceptable state, potential areas of improvement were identified.

From a theoretical perspective, the cognitive subtests measured Gc to a great extent, Gf to some extent, and other types of intelligence to a limited extent or not at all. From a competency mapping perspective, the findings were mixed, but the only clear observation was that WK did not map strongly to the officer competencies across independent efforts and does not contribute to the officership or aircrew composites.

Considering the results of the aforementioned evaluations, we recommend that AFOQT Form U developers consider several revisions to its content. These include (1) not bringing back any decommissioned subtests; (2) excluding WK due to its poor psychometric performance and poor linkage to the USAF officer competencies; (3) modifying some operational subtests; and (4) expanding the assessment of cognitive ability to address theoretical deficiencies (e.g., more focus on Gf), improve linkages with critical competencies, and reduce SGDs. We are confident that these enhancements would improve the psychometric integrity of the AFOQT as a whole. See Table 8.

Table 8. AFOQT Form U Recommendations

Cognitive Subtests to Include with Slight Modifications	Cognitive Subtests to Include with Substantial Modifications	Cognitive Subtests to Exclude/Not Consider
AR, RC, TR, IC	VA, MK, PS, BC, AI	WK, EM, GS, HF, MC, DI, RB, SR

Note. Underlined subtests appear in Form T.

There are several limitations to these findings. The first and main limitation is that the empirical evaluation of the cognitive subtests was based solely on the metrics included in the publicly available technical reports. Because the foci of the AFOQT assessments have been changing over the years, the extent of the psychometric analyses performed on the subtests and their reporting also changed. For example, since the publication of the Air Force Diversity and Inclusion initiative, the scope of the analyses placed more emphasis on mean score SGDs, whereas in the past years the scope did not always include the reporting of the effect sizes. The second limitation is that most recent needs analyses and mapping tasks were performed for the operational cognitive subtests (i.e., AFOQT Form T) only, which did not consider discontinued subtests from earlier forms (Kantrowitz et al., 2022). This leaves us wondering if the de-commissioned subtests would have mapped to the attributes and competencies for the aircrew career fields.

6.0 CONCLUSION

The present research was sparked by the USAF's continuous effort to improve the AFOQT. The goals for Form U are to (1) minimize mean score SGDs for gender, race, and ethnic minorities and (2) maximize prediction of officer-critical performance in training and beyond. These goals have been a focus of the last 100 years of study in the field of I/O psychology. Our three-pronged technical approach sought to balance the empirical, theoretical, and competency evaluation considerations.

Based on the findings of our evaluation, we predict that the suggested enhancements to the AFOQT would (1) capitalize on the psychometric strengths of the subtests while mitigating their weaknesses; (2) align the assessment with the dominant theoretical frameworks which would ensure the right amount of measurement precision and competency mapping; and (3) target critical officer competencies which would ensure predictive validity.

7.0 REFERENCES

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied*, *6*(4), 259.

Ackerman, P. L., Beier, M. E., & Boyle, M. D. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of experimental psychology: General, 131*(4), 567.

Aguinis, H., Henle, C. A., & Ostroff, C. (2001). Measurement in work and organizational psychology. *Handbook of industrial, work, and organizational psychology*, *1*, 27-50, p. 7

Carretta, T. R., Rose, M. R., & Trent, J. D. (2016). *Air Force Officer Qualifying Test Form T: Initial item-, test-, factor-, and composite-level analyses,* AFRL-RH-WP-TR-2016-0093. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Warfighter Interface Division.

Carretta, T. R. (1987). *Basic Attributes Test (BAT) system: Development of an automated test battery for pilot selection, 9* AFHRL-TR-87-9. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower, and Personnel Division.

Carretta, T. R. (1990). Cross validation of experimental USAF pilot training performance models. *Military Psychology*, *2*(4), 257-264.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, *38*, 592.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*(3), 153–193.

Cattell, R. B. (1950). *Personality: A systematic theoretical and factual study*. McGraw-Hill.

Corsini, R. (1957). *Methods of group psychotherapy*.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98-101.

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

Cucina, J. M., & Howardson, G. N. (2017). Woodcock-Johnson-III, Kaufman Adolescent and Adult Intelligence Test (KAIT), Kaufman Assessment Battery for Children (KABC), and Differential Ability Scales (DAS) support Carroll but not Cattell-Horn. *Psychological Assessment*, *29*(8), 1001.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological methods*, *2*(3), 292.

Drasgow, F., Nye, C. D., Carretta, T. R., & Ree, M. J. (2010). Factor structure of the Air Force Officer Qualifying Test Form S: Analysis and comparison with previous forms. *Military Psychology*, *22*(1), 68-85.

Furr, R. M., & Bacharach, V. R. (2008). Psychometrics and the importance of psychological measurement. *Psychometrics. Thousand Oaks, CA: Sage Publications Inc*.

Jensen, A. R. (1998). *The g factor*. Westport, CT: Praeger.

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. *International Journal of Aerospace Psychology, 3-4,* 109-120.

Kantrowitz, T., Segall, D., Kingry, D., Valone, A., Mann, K., Walsh, J., Wilson, R., Trent, J., & Carretta, T. (2023). *Air Force Officer Qualifying Test (AFOQT) Form T Version 2,* AFRL-RH-WP-TR-2023-0021. Wright-Patterson AFB, OH: 711 Branch.

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,* AFRH-RH-WP-TR-2022-0038. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

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.

Lang, J. W. B., Kersting, M., Hulsheger, U. R., & Lang, J. (2010). General mental ability, narrower cognitive abilities, and job performance: The perspective of nested-factors model of cognitive abilities. *Personnel Psychology, 63,* 595-640.

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

Makransky, G., & Glas, C. A. (2011). Unproctored internet test verification: Using adaptive confirmation testing. *Organizational Research Methods*, *14*(4), 608-630.

Mann, K. J., Drake, M. R., LeRoy, V. H., Carretta, T. R., Mouton, A. N., & Deregla, A. R. (2023). Self-Description Inventory Plus (SDI+): Item-, facet-, & domain-level analyses, AFRL-RH-WP-TR-2023-0066. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In *Handbook of modern item response theory* (pp. 101-121). NY: Springer.

McDaniel, M. A., & Banks, G. C. (2010). Cognitive ability. In D. H. Reynolds & J. C. Scott (Eds.), *The Handbook of Workplace Assessment: Selecting and Developing Organizational Talent.* CA: Pfeiffer.

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

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17,* 609-626.

Persing, C.R., Woolley, M. R., Sizemore, S. J., Dirr, B.A., & Carretta, T.R. (in press). *Development of criterion measures related to Air Force competencies and attributes*, AFRL-RH-WP-TR-2023-xxxx. Wright-Patterson AFB, OH: 711 Human Performance Wing, Air and Space Biosciences Division, Performance Optimization Branch.

Parish, C., Morath, R., Lodato, M., & Stachowski, A. (2008). *AFOQT Form S effectiveness analyses: Composite-level analyses and report.* Randolph AFB, TX: Air Force Personnel Center, Strategic Research and Assessment Branch.

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.

Portman-Tiller, C. A., Biggerstaff, S., & Blower, D. (1998). *Relationship between the aviation selection test and a psychomotor battery*. Naval Aerospace Medical Research Laboratory, Pensacola, FL.

Raven, J. C., & Court, J. H. (1998). *Raven's progressive matrices and vocabulary scales* (pp. 223-237). Oxford: Oxford Psychologists Press.

Ree, M. J., & Carretta, T. R. (2022). Thirty years of research on general and specific abilities: Still not much more than g. *Intelligence*, *91*, 101617.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span

tasks. *European Journal of Psychological Assessment*.
Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of personality assessment*, *81*(2), 93-103.
Reise, S. P., & Waller, N. G. (2002). *Item response theory for dichotomous assessment data*.
Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86,* 162-173.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological assessment*, *8*(4), 350.
Schneider, J. W., & Newman, D. A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review, 25,* 12-27.
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*. Randolph AFB, TX: Air Force Personnel Center.
Sizemore, S. J., Wilson, R., Mann, K. J., Carretta, T. R., Mouton, A. N., Deregla, A. R., & Infoscitex. (2022). *Air Force Officer Qualifying Test (AFOQT): Situational Judgment Test (SJT) revision,* AFRL-RH-WP-TR-2022-0076. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA* (Vol. 724). Belmont, CA: Thomson/Brooks/Cole.
Unsworth, N., Spillers, G. J., & Brewer, G. A. (2009). Examining the relations among working memory capacity, attention control, and fluid intelligence from a dual-component framework. *Psychological Test and Assessment Modeling, 51*(4), 388.
van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). Dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113,* 842-861.
Vangent. (1993). *Computer Programmer Aptitude Battery examiner's manual*. Chicago, IL: Author.
Walsh, J. L., Brady, M. F., Woolley, M. R., & Carretta, T. R. (2022). *Air Force Officer Qualifying Test (AFOQT) Form T evaluation: Item-level analyses*, AFRL-RH-WP-TR-2022-0037. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
Walsh, J. L., Woolley, M. R., Brady, M. F., Melick, S. R., & Carretta, T. R. (2021). *Air Force Officer Qualifying Test (AFOQT) Form T: Psychometric evaluation of the Situational Judgment Test*, AFRL-RH-WP-TR-2021-0080. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
Woolley, M. R., Walsh, J. L., Mann, K. J., Wilson, R. T., Carretta, T. R., Mouton, A. N., & Deregla, A. R. (2022). *Self-Description Inventory-Officer (SDI-O): Item-, facet-, and domain-level analyses* (p. 414). AFRL-RH-WP-TR-2022-0091. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
Wechsler, D. (1997). *Wechsler Memory Scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.
Zickar, M. J. (2012). A review of recent advances in item response theory. *Research in personnel and human resources management*, *31*, 145-176.
Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical

test theory and item response theory. In C. E. Lance & R. J. Vanderbergh (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 37-59). New York, NY: Routledge.

APPENDIX A – Brief Overview of Classical Test Theory and Item Response Theory

Classical Test Theory (CTT)

CTT is a commonly used and widely accepted item-level analysis methodology (Furr & Bacharach, 2008; Nunnally & Bernstein, 1994). CTT rests on several assumptions, including that (1) the true and error scores are uncorrelated; (2) the average error score across replications for each test-taker is zero; and (3) the error scores on parallel tests are uncorrelated (Zickar & Broadfoot, 2009). Among CTT's advantages is the ability to use smaller sample sizes, relative computational ease of calculating the statistics, and the ability to work with multidimensional data (Zickar & Broadfoot, 2009). However, CTT is frequently criticized because its statistics and parameters are sample- and test-dependent which makes the comparability among the test administrations difficult. Secondly, test-takers' and item parameters are not on the same scale. Finally, reliability is the estimate of the whole test assuming equal error variance which is hard to satisfy in the real world (Zickar & Broadfoot, 2009). Due to its shortcomings, CTT is often supplemented and more frequently completely replaced with IRT. The areas of CTT investigation are described in more detail below.

Descriptives

The first step in this research involved the examination of the item descriptive statistics, including means (*M*), *SD*, skewness, and kurtosis. These descriptives, also known as measures of central tendency and variability, help identify poor performing and well performing items. As Kline (2005) states, "Generally, the higher the variability of the items and the more the mean of the item is at the center point of the distribution, the better the item will perform" (p. 95). Recall that the items were scored dichotomously and therefore the items' *M* were also the items' difficulty parameters (*p*-values, described in more detail below). Negative skewness indicates that the majority of the scores are clustered around the upper end of the distribution and therefore mark an easier item; conversely, positive skewness indicates that the majority of the scores are clustered around the lower end of the distribution and therefore mark a more difficult item (Tabachnick & Fidell, 2007). The skewness of 0 indicates a normal distribution; the skewness $\leq$ -3 or $\geq$ 3 is considered extreme (Kline, 2005). Positive kurtosis indicates heavy tails and peakedness relative to the normal distribution, whereas negative kurtosis indicates light tails and flatness (DeCarlo, 1997). Most studies do not interpret kurtosis, and some do not even report it. With the larger sample sizes such as in this study, the impact of negative or positive kurtosis diminishes (Tabachnick & Fidell, 2007).

Item Difficulty

Item difficulty was assessed by observing *p*-values, which represent the percentage of test-takers who endorsed the correct response option. As mentioned above, *p*-values also represent the item's mean (*M*; Kline, 2005). High *p*-values indicate an easy item and low *p*-values indicate a difficult item (Kline, 2005). In general, items that have difficulty of 0 or 1 are useless because they do not provide variability (information) about a test-taker. More specifically, items with *p*-values $\leq$ .20 for all subtests with five response options ($\leq$ .25 for IC which has four response options) are considered too difficult and represent a floor effect. Items with five response options that have *p*-values $\geq$ .80 ($\geq$ .75 for IC) are considered too easy and represent a ceiling effect. Literature suggests that items with *p*-values of .50 are optimal, because they provide the best differentiation between low-ability and high-ability test-takers (Kline, 2005). However, the evaluation of whether an item

is difficult or easy should be based on the goals of the selection instrument. Thus, we used slightly different cutoffs based on our dataset and goals. Specifically, our dataset *p*-values ranged from .38 to .73, rendering cutoffs below .25 and above .75 useless. Instead, we used the following cutoffs: .60-.73 too easy, .50-.60 moderate/acceptable difficulty, and .38-.50 good/high difficulty.

Item Discriminability

Item discriminability, expressed as a corrected ITC, refers to a correlation between each item and the total subtest score (computed with the item in question removed; Kline, 2005). Much like the Pearson product moment correlation, ITC's range from -1 to 1, where values closer to 1 are considered strong and values closer to 0 are considered weak. Negative values indicate that the item is negatively related to the subtest score, which is highly undesirable.

In our dataset ITCs ranged from .27 to .68. Due to a truncated range of values, we created evaluation criteria that are more useful than recommendations typically in the literature, which might be more useful with a larger range of values. Our evaluation criteria for item difficulty are as follows: 0-.30 as poor, .31-.45 as acceptable, and greater than .45 as good.

Internal Consistency

Internal consistency at the subtest level, expressed as Cronbach's alpha, concerns the interrelatedness of items within a subtest (Schmitt, 1996). The literature typically recommends cutoffs of .60 as acceptable for research purposes, .70 as acceptable for practice, .80 as desirable, and .90 and above as very good (Cronbach, 1951). Internal consistencies in our dataset ranged from .66 to .91. Based on this, we used the criteria of .70-.80 as acceptable, .80-.90 as good, and below .70 or above .90 as poor/unacceptable. This is because above .90 might indicate redundancy in the subtest, and below .70 is not acceptable in a high-stake testing situation. Note that Cronbach alpha is not an appropriate metric for speeded tests, such as TR and BC.

Item Mean Score Subgroup Differences

For each subtest we examined mean score differences across legally protected subgroups between the majority subgroup (e.g., White, males) and minority subgroups (e.g., gender, racial, and ethnic subgroups). The subgroups examined in this effort were – Female/Male (F/M), B/W, Asian/White (A/W), and Hispanic/Non-Hispanic (H/nH). These differences are expressed using Cohen's *d*, which is a measure of mean score difference in SD units (Cohen, 1992). Cutoffs of |.40| and |.80| were chosen to represent moderately and highly problematic subgroup differences, respectively. Although these cutoffs differ slightly from the ones prescribed by the literature, they are in line with the cutoffs used for the Armed Services Vocational Aptitude Battery (ASVAB) and other standardized assessments throughout the Department of Defense (DoD). Moderate to large effect sizes suggest that an item may inadvertently favor one subgroup over the other and therefore may contribute to adverse impact.

Criterion Validity

The final criterion we used to evaluate the AFOQT subtests was their relative criterion-related validity. There are no recommended cutoffs for criterion-related validity values in the literature, so our evaluation was based on relative prediction compared to the other AFOQT subtests in the dataset. It is important to note that criteria changed over the years, so while we can compare validity coefficients, they are not a direct comparison across Forms.

Item Response Theory (IRT)

IRT is a family of mathematical models that estimate the probability of test-takers' endorsing different response options as a function of their standing on the latent trait (Lord & Novick, 2008). In the areas of psychological measurement, IRT models are usually applied to data (1) scored dichotomously (e.g., correct answers scored as '1' and incorrect answers scored as '0'; 1PL, 2PL, 3PL) or polytomously (e.g., using common Likert-type scales; Generalized Partial Credit Model [GPCM], Graded Response Model [GRM]); and (2) tapping a single construct (unidimensionality) or multiple constructs (multidimensionality).

In applying IRT to the AFOQT cognitive subtests, it is best to use dichotomous IRT models (1PL, 2PL, and 3PL) because (1) the items are scored as '0' and '1' and (2) the results of the Principal Components Analyses (PCA) typically reveal sufficient unidimensionality for each subtest. 1PL model generates one parameter (difficulty); 2PL model generates two parameters (difficulty and discriminability); and 3PL model generates three parameters (difficulty, discriminability, and pseudo-guessing). For a general information on the IRT models and frameworks, please refer to Lord and Novick (2008), Reise and Henson (2003), or Zickar (2012).

b Parameter

Difficulty parameter (*b*) describes a point on the latent trait continuum where the probability for endorsing a correct response .50. For example, a *b* parameter equal to -1.29 for VA Item T1-01 would indicate that test-takers with a latent trait standing lower than -1.29 would likely get the item wrong; test-takers with a latent trait standing higher than -1.29 would get the item right; and test-takers with a latent trait of exactly -1.29 would have a 50% chance of getting the item right. Depending on the goals of a psychological instrument, it might be desired to target certain ranges of the latent trait. Therefore, the items with the *b* parameter within the desired range should provide the most amount of information. For example, if the desire to filter out test-takers with lower standing on verbal ability, items that make up a subtest designed to measure the construct should target latent trait range within two SDs below the mean (the items would be fairly easy). Applying the literature recommendation to the AFOQT (select-out strategy), *b* parameters below two SD of the mean would be considered easy; within two SD of the mean – average; and above mean – difficult.

a Parameter

Discriminability parameter (*a*) describes the extent to which the item is capable of differentiating between test-takers with lower and higher standing on the latent trait (Masters & Wright, 1997). Generally, *a* parameters below 1.00 are considered as indicators of poor items and *a* parameters above or at 1.00 are considered as indicators of good items (Reise & Waller, 2002).

c Parameters

Pseudo-guessing parameter (*c*) reflects the probability that an a low-ability test-taker endorses a correct response on a cognitive ability test by guessing. Guessing may fluctuate due to factors beyond pure guessing (some distractors may be obviously wrong; Zickar, 2012). Evidence exists that guessing parameter adds little to the precision of theta estimate (Makransky & Glas, 2011). Note that the AFOQT test-takers are not penalized for guessing.

APPENDIX B – Psychometric Cutoffs Used in the Current Effort

Metric	Cutoffs Used for Evaluation	Literature Recommendations
p-value (difficulty)	Low (red): .60 - .70 Medium (orange): .50 - .60 High (green): .40 - .50	Kline (2005) recommends cutoffs of .20 being too difficult and .80 being too easy to be useful.
ITC (discriminability)	Poor (red): < .30 Acceptable (orange): .30 - .45 Good (green): > .45	"Item-total correlations above .30 are preferred" (Aguinis et al., 2001)
Cronbach's alpha (internal consistency)	Poor (red): < .70 and > .90 Acceptable (orange): .70 - .80 Good (green): .80 - .90	Cronbach (1951) and Schmitt (1996) recommend the cutoffs we used.
Cohen's *d* (subgroup mean score differences)	Adverse Impact Highly Likely (red): > .80 Adverse Impact Likely (orange): .40 - .80 Adverse Impact Unlikely (green): < .40	Shore et al., (2020) notes USAF interprets *d* = .40 as indicating adverse impact. Traditional cutoffs define <.20 as small, .50 as moderate, and .80 or greater as large.
Criterion-related validity	Criterion-related validity evaluations are relative to other subtests, and thus there are no evaluation cutoffs. Best relative predictors are in green Average relative predictors are in orange Poor relative predictors are in red	N/A

Note. N/A = Not Applicable; ITC = Item Total Correlations.

APPENDIX C – Theoretical Frameworks Reviewed in the Current Effort

Cattell's Crystallized and Fluid (C&F) Intelligence Model	
Form of Intelligence	Definition
Fluid (Gf)	Ability to solve novel problems through reasoning
Crystallized (Gc)	Reliance on prior knowledge and experiences to solve problems

Cattell-Horn-Carroll (CHC) Theory of Intelligence	
Form of Intelligence	Definition
Comprehension Knowledge	Language comprehension and general knowledge
Domain-Specific Knowledge	Declarative and procedural knowledge of specialized interests
Reading/Writing	Declarative and procedural knowledge related to literacy
Quantitative Knowledge	Declarative and procedural knowledge related to mathematics
Fluid Reasoning	Use deliberate and controlled mental operations to solve novel problems
Short-Term Memory	Apprehend and maintain awareness of information that is useful for multi-step problem-solving
Long-Term Storage & Retrieval	Store and consolidate new information and fluently retrieve stored information
Processing Speed	Automatically and fluently perform relatively easy elementary cognitive tasks
Reaction & Decision Speed	Speed at which very simple perceptual discriminations are performed
Visual-Spatial Processing	Perceive, discriminate, and manipulate images
Auditory Processing	Perceive, discriminate, and manipulate sounds
Olfactory Processing	Perceive, discriminate, and manipulate smells
Kinesthetic Processing	Perceive, discriminate, and manipulate sensations of body movements
Tactile (Haptic) Processing	Perceive, discriminate, and manipulate touch stimuli
Psychomotor abilities	Skilled performance of motor tasks
Psychomotor Speed	Speed of motor functions

Thurstone's Primary Mental Abilities (PMA) Model	
Form of Intelligence	Description
Word Fluency	Ability to use words, fluency in use of words
Verbal Comprehension	Ability to understand words, concepts, and ideas
Numerical Ability	Ability of computation using numbers
Spatial-Visualization	Ability to visualize and manipulate patterns and forms in 3-D space
Perceptual Speed	Ability to understand perceptual details quickly and effectively
Memory	Ability to recall information effectively
Inductive Reasoning	Ability to derive general components and rules from presented information

APPENDIX D – Unmeasured or Deficiently Measured CHC Types of Intelligence and Sample Assessments

Based on the mapping between the AFOQT cognitive subtests and the prevailing theoretical frameworks of intelligence, several competencies have been identified as important for all USAF officer candidates and for AFSC-specific candidates. These competencies are not currently measured or are measured deficiently by the AFOQT subtests. The table below lists these competencies with samples of assessment methods.

Types of CHC Intelligence	Sample Assessments
Fluid Intelligence (deficient)	• Raven's Advanced Progressive Matrices (Raven & Court, 1998) • Cattell's Culture Fair Intelligence Test (CFIT)
Inductive reasoning/fluid reasoning (deficient)	• Non-Verbal Reasoning Test (Corsini, 1957; Klein et al., 2015) • Speeded Letter Series Test (Klein et al., 2015; Vangent, 1993)
Memory/controlled attention (unmeasured)	• Complex Span Tests (Redick et al., 2012; Unsworth et al., 2009) • Wechsler Memory Scale III Test (Wechsler, 1997)
Reaction speed (unmeasured)	• Coding Speed (ASVAB)
Perceptual speed (unmeasured)	• Processing Speed – Scanning Tests (Ackerman & Cianciolo, 2000; Ackerman et al., 2002)
Psychomotor skills (unmeasured)	• Basic Attributes Test (BAT; c.f., Carretta, 1987, 1990) • Computer Based Performance Test (CBPT; Portman-Tiller et al., 1998)

Note. Some tests listed above may involve components that stray from suggested competencies. Focus should be placed on the competencies noted on the left column when examining sample assessments.

APPENDIX E – Competency Mapping Evaluation

Subtest	Number of Core Officer Competencies with Moderate Linkages	Number of Core Officer Competencies with Strong Linkages
VA	2	0
AR	2	0
WK	0	0
MK	1	0
RC	1	0
PS	0	0
TR	1	0
IC	2	0
BC	1	0
AI	0	0
DI	1	1
MC	0	0
EM	2	0
RB	1	1
HF	0	0
GS	0	0
SR	1	0

Note. In green are subtests that strongly linked to competencies/attributes; in orange are subtests that moderately linked to competencies/attributes; in red are subtests that did linked either poorly or not at all to competencies/attributes. Underlined subtests appear on Form T.

Subtest	Number of AFSC-Specific Competencies with Moderate Linkages	Number of AFSC-Specific Competencies with Strong Linkages
VA	2	0
AR	0	0
WK	2	0
MK	0	0
RC	0	0
PS	0	0
TR	1	0
IC	1	0
BC	0	0
AI	0	0
DI	0	0
MC	0	0
EM	1	0
RB	0	0
HF	0	0
GS	0	0
SR	1	0

Note. In orange are subtests that moderately linked to competencies/attributes; in red are subtests that did linked either poorly or not at all to competencies/attributes. Underlined subtests appear on Form T.

Subtest	Number of Rated Career Field Competencies with Moderate Linkages*	Number of Rated Career Field Competencies with Strong Linkages*
VA	2	1
AR	3	5
WK	3	0
MK	3	3
RC	3	2
PS	1	2
TR	2	4
IC	3	7
BC	4	3
AI	3	2

Note. In green are subtests that strongly linked to competencies/attribute; in orange are subtests that moderately linked to competencies/attributes. Underlined subtests appear on Form T.

Subtest	Number of AFOQT Form T Composites	List of the AFOQT Form T Composites
VA	0	[Did not make it]
AR	1	Academic Aptitude
WK	0	[Did not make it]
MK	2	• ABM • CSO
RC	1	General Officership
PS	1	Academic Aptitude
TR	1	Academic Aptitude
IC	2	• Pilot • CSO
BC	0	[Did not make it]
AI	3	• Pilot • CSO • ABM

Note. In green are subtests that survived the analyses; in red are subtests that did not survive the analyses. Underlined subtests appear on Form T.

APPENDIX F – Unmeasured or Deficiently Measured Competencies and Sample Assessments

Based on the competency mapping evaluation, several competencies have been identified as important for all USAF officer candidates and specific officer occupations. These competencies are not currently measured by cognitive AFOQT subtests (however, as the table shows, they may be measured by the non-cognitive subtests such as SDI-O and SJT). The table below lists these competencies with examples of potential assessment methods.

Core and AFSC-Specific Officer Competencies	Example Assessments
Accepts Feedback	• SJT (or higher fidelity sim) • Personality • Biodata
Accountability	• Personality • Biodata
Communication: Active Listening	• Structured Interview • Biodata
Followership	• SJT (or higher fidelity sim)
Initiative	• Personality • Biodata
Integrity	• Personality • Biodata • Integrity Scales
Perseverance	• Personality • Biodata
Professionalism	• Personality
Self-Awareness	• Personality • Biodata • Structured Interview • SJT (or higher fidelity sim)
Collaboration	• Personality • SJT (or higher fidelity sim)
Continuous Learner	• Personality
Cultural Awareness	• Personality • SJT (or higher fidelity sim)
Leadership: Vision & Influence	• Personality • SJT (or higher fidelity sim) • Structured Interview
Leadership: Strategic Thinking	• Personality • SJT (or higher fidelity sim)

	• Structured Interview
Leadership: Fostering Innovation	• Personality • SJT (or higher fidelity sim) • Structured Interview
Openness to Alternative Views	• Personality • Biodata
Resilience	• Personality • Biodata
Self-Control	• Personality
Takes Care of People	• SJT (or higher fidelity sim) • Personality

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

1PL	1-parameter logistic model in IRT
2PL	2-parameter logistic model in IRT
3PL	3-parameter logistic model in IRT
a parameter	A discriminability parameter in IRT
ABM	Air Battle Manager
AECP	Airman Education and Commissioning Program
AFOQT	Air Force Officer Qualifying Test
AFPC/DSYX	Air Force Personnel Center Strategic Research and Assessments Branch
AFPTP	Air Force Personnel Testing Program
AFROTC	Air Force Reserve Officer Training Corps
AFSC	Air Force Specialty Code
AI	Aviation Information
AR	Arithmetic Reasoning
ASVAB	Armed Services Vocational Aptitude Battery
A/W	Asian/White
b parameter	A difficulty parameter in IRT
B/W	Black/White
BC	Block Counting
c parameter	A guessing parameter in IRT
C&F	Crystalized & Fluid (intelligence)
CHC	Cattell-Horn-Carroll's (model of intelligence)
CSO	Combat Systems Officer
CTT	Classical Test Theory
DI	Data Interpretation
DTIC	Defense Technical Information Center
EM	Electrical Maze
F/M	Female/Male
g	The general intelligence factor
Gc	Crystalized Intelligence
Gf	Fluid Intelligence
GPA	Grade Point Average
GS	General Science
H/nH	Hispanic/Non-Hispanic
HF	Hidden figures
I/O	Industrial-Organizational psychology
IC	Instrument Comprehension
IRT	Item Response Theory
ITC	Item-Total Correlation
M	Mean
MC	Mechanical Comprehension
MK	Math Knowledge
N/A	Not Applicable
OTS	Officer Training School

PMA	Primary Mental Abilities
PS	Physical Science
p-value	Proportion of correct items compared to all items in the assessment in CTT
RB	Rotated Blocks
RC	Reading Comprehension
RPA	Remotely Piloted Aircraft
SD	Standard Deviation
SDI+	Self-Description Inventory Plus
SDI-O	Self-Description Inventory - Officers
SGD	Subgroup Differences
SJT	Situational Judgment Test
SR	Scale Reading
TR	Table Reading
UPT	Undergraduate Pilot Training
URT	Undergraduate RPA Training
USAF	United States Air Force
VA	Verbal Analogies
WK	Word Knowledge
≤	Less than or equal to
≥	Greater than or equal to