



AFRL-RH-WP-TR-2023-0021

Development of the Air Force Officer Qualifying Test (AFOQT) Form T Version 2

**Tracy Kantrowitz Daniel Segall
Donna Kingry Amanda Valone
Personnel Decisions Research Institutes**

**Kyle Mann Julia Walsh
Rusty Wilson
Infoscitex**

**John Trent
Air Force Personnel Center**

**Thomas Carretta
Air Force Research Laboratory**

**June 2023
Interim Report**

DISTRIBUTION STATEMENT A. Approved for public release.

**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-2023-0021 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION.

CARRETTA.T
HOMAS.R.122
8984929

Digitally signed by
CARRETTA.THOMAS.R.
1228984929
Date: 2023.06.15
11:23:24 -04'00'

THOMAS R. CARRETTA, PhD
Work Unit Manager
Performance Optimization Branch
Airman Biosciences Division

WILLIAMS.LO
GAN.ANDREW
.1273597634

Digitally signed by
WILLIAMS.LOGAN.ANDR
EW.1273597634
Date: 2023.06.15 15:56:25
-04'00'

LOGAN R. WILLIAMS, DR-III, PhD
Human Performance Area Lead
Operational Product Section
Product Development Branch
Airman 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	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YY) 01-06-23		2. REPORT TYPE Interim		3. DATES COVERED (From - To) 9/14/21 – 2/28/23	
2. TITLE AND SUBTITLE Development of the Air Force Officer Qualifying Test (AFOQT) Form T Version 2				5a. CONTRACT NUMBER FA8650-21-F-4104	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER PE62202F	
6. AUTHOR(S) Tracy Kantrowitz ^a , Daniel Segall ^a , Donna Kingry ^a , Amanda Valone ^a , Kyle Mann ^b , Julia Walsh ^b , Rusty Wilson ^b , John Trent ^c , and Thomas Carretta ^d				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER H12Q	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Personnel Decisions Research Institutes ^a 1911 N. Fort Myer Drive Suite 410 Arlington, VA 22209		Infoscitex Corporation ^b 4027 Colonel Glen Hwy Suite 210 Dayton, OH 45431-1672		8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command ^d Air Force Research Laboratory 711 th Human Performance Wing Airman Systems Directorate Airman Biosciences Division Performance Optimization Branch Wright-Patterson AFB, OH 45433		Air Force Personnel Center ^c 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-2023-0021	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution Statement A: Approved for public release.					
13. SUPPLEMENTARY NOTES Report contains color. AFRL-2023-2913, cleared 15 June 2023					
14. ABSTRACT This technical report describes the development of the Air Force Officer Qualifying Test (AFOQT) Form T Version 2 (Form Tv2). The purpose of this project was to develop and field a replacement of AFOQT Form T Version 1 (Form Tv1). The focus was on replacing problematic items based on previous analyses, reducing mean score differences for underrepresented subgroups, and implementing the assessment for web-based administration. This report is divided into the following sections: (1) Executive Summary, (2) Introduction, (3) Method, Assumptions, and Procedures, (4) Results and Discussion, (5) Conclusion, and (6) References. Each section provides detailed results and descriptions of processes used to develop and field the AFOQT Form Tv2. Across the objectives, the results indicated that, compared to Form Tv1, Form Tv2 yielded similar levels of score precision and increased potential for diversity.					
15. SUBJECT TERMS Air Force Officer Qualifying Test, AFOQT, personnel assessment, classification					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT: SAR	18. NUMBER OF PAGES 66	19a. NAME OF RESPONSIBLE PERSON (Monitor) Thomas R. Carretta 19b. TELEPHONE NUMBER (Include Area Code)
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			

TABLE OF CONTENTS

ACKNOWLEDGEMENT	iv
1.0 Executive Summary	1
2.0 INTRODUCTION	3
3.0 METHOD, ASSUMPTION AND PROCEDURES	5
3.1 Objective 1: Replacement of Problematic AFOQT Form Tv1 Items for Current Subtests	5
3.2 Objective 2: Data Collection, Item Analysis, and Scale Development ..	6
3.2.1 Identification of Anchor Items for Inclusion in Data Collection	6
3.2.2 Development of a Data Collection Strategy	6
3.2.3 Preparing Test Materials	9
3.2.4 Data Collection: AFOQT Form Tv2 A and B	10
3.2.5 Item Analysis Procedures	11
3.2.6 Form Assembly Goals and Process.....	12
3.3 Objective 3: Equivalency/Equating Study	13
3.4 Objective 4: Develop Two Forms of AFOQT Form Tv2 and Support New Composite Scoring	13
4.0 RESULTS AND DISCUSSIONS.....	14
4.1 Objective 1: Replacement of Problematic AFOQT Form Tv1 Items for Current Subtests	14
4.2 Objective 2: Data Collection, Item Analysis, and Scale Development	15
4.2.1 Data Cleaning.....	15
4.2.2 IRT Analysis	16
4.2.3 Form Assembly Results	16
4.2.4 Diversity Index Results.....	18
4.2.5 Reliability Estimates	24
4.2.6 Test Information Function (TIF).....	24
4.2.7 Subtest Composition	27
4.3 Objective 3: Equivalency/Equating Study	27
4.3.1 Equating Tables	27
4.3.2 Equating Approach for TR.....	30
4.4 Objective 4: Develop Two Forms of AFOQT Form Tv2 and Support New Composite Scoring	31
4.4.1 Composite Equating Results	31
4.5 Discussion.....	45
5.0 CONCLUSION.....	45
6.0 REFERENCES	47
Appendix. Classical Test Theory item-level analyses	49
LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS.....	58

LIST OF TABLES

Table 1. AFOQT Form Tv1 Subtest Descriptions	3
Table 2. Details of Data Collection for AFOQT Form Tv2 (A and B)	11
Table 3. Counts for Items Developed and Pilot Tested	14
Table 4. Sample Sizes after Data Cleaning.....	16
Table 5. Latent Ability Distribution Average Values based on Form Tv1 and Tv2 Analysis	17
Table 6. Diversity Index Results for Race and Gender - Form T1, Form Tv2 (A), Form T2, Form Tv2 (B), and Increase in Diversity for Forms Tv2 (A and B).....	19
Table 7. Diversity Index Results by Subgroup Race/Ethnicity – Tv2 (A) and (B) and Changes Relative to Form T1 and T2.....	20
Table 8. Coefficient Alpha Reliability Estimates for Form Tv1 and Form Tv2 Subtests	24
Table 9. TIF	25
Table 10. Final Item Counts by Source for Form Tv2.....	27
Table 11. Form Tv2(A) Number Right (NR) Converted to Equated <i>t</i> -Scores.....	28
Table 12. Form Tv2(B) Number Right (NR) Converted to Equated <i>t</i> -Scores.....	29
Table 13. Sample Sizes associated with TR Data Collection	30
Table 14. Rate Score Comparisons between OTS and SOS Subsamples – Web-Based Administration.....	30
Table 15. Number Right Score Comparisons between OTS and SOS Subsamples – Paper-Based Administration	31
Table 16. Correlation and Reduction in Uncertainty Values between Form Tv1 and Form Tv2 Composites	32
Table 17. Composite Reliability Estimates for Form Tv1 and Form Tv2	32
Table 18. Criterion-Related Validity Estimates for Form Tv1 and Form Tv2 Composites.....	33
Table 19. Subgroup Differences and Passing Rate Comparison – Pilot Composite	34
Table 20. Subgroup Differences and Passing Rate Comparison – CSO Composite	34
Table 21. Subgroup Differences and Passing Rate Comparison – ABM Composite	35
Table 22. Faking Evaluation of SDI-O Facets and Effects on Composite Score Distributions.....	36

LIST OF FIGURES

Figure 1. Distribution of Score Differences for Form Tv1 and Form Tv2 Composites – Pilot.....	37
Figure 2. Cumulative Score Difference Density for Form Tv1 and Form Tv2 Composites – Pilot.....	38
Figure 3. Distribution of the Score Difference Between Maximum Composite and Form Tv1 Composite – Pilot.....	39
Figure 4. Cumulative Score Difference Density for Maximum Composite Score and Form Tv1 Composite Score – Pilot.....	39
Figure 5. Distribution of Score Differences for Form Tv1 and Form Tv2 Composites – CSO.....	40
Figure 6. Cumulative Score Difference Density for Form Tv1 and Form Tv2 Composites – CSO.....	40
Figure 7. Distribution of Score Difference for Maximum and Form Tv1 Composite – CSO	41
Figure 8. Cumulative Score Difference Density for Maximum and Form Tv1 Composite Score – CSO	42
Figure 9. Distribution of Score Differences for Form Tv1 and Form Tv2 Composites – ABM.....	43
Figure 10. Cumulative Score Difference Density for Form Tv1 and Form Tv2 Composites – ABM.....	43
Figure 12. Cumulative Score Difference Density for Maximum and Form Tv1 Composite Score – ABM.....	44

ACKNOWLEDGEMENT

We would like to sincerely thank the following individuals for their valuable contributions to this effort:

Sophie Romy, PhD (Air Force Personnel Center)
Captain Steven Zimmer (Air Force Personnel Center)
Captain Zachary Baumann (Air Force Personnel Center)
Montana Woolley, PhD (Infoscitex)
Annette Rizer (Infoscitex)

1.0 EXECUTIVE SUMMARY

The recruitment, selection, development, and retention of a diverse workforce has been an important objective for the United States military since the publication of the Diversity and Inclusion (D&I) strategic plan in 2012. For the United States Air Force (USAF), recruiting and identifying officer candidates is of critical importance to ensuring a qualified and diverse workforce. This current project focuses on addressing the D&I initiative in commissioning and classification of officers through the use of the Air Force Officer Qualifying Test (AFOQT), a critical component of the Air Force Personnel Testing Program (AFPTP) in use since 1953. The AFOQT has been considered as part of the USAF's D&I initiatives because some selection procedures that strongly predict training and/or job performance (which the AFOQT has historically demonstrated) also show systematic differences in mean scores across demographic groups such as race, ethnicity, and gender (Bobko & Roth, 2013; Ployhart & Holtz, 2008), which hinder the goals of D&I by selecting minority applicants at lower rates than majority applicants.

The AFOQT has served as a robust tool for officer commissioning and for classification of officers into rated (i.e., aircrew) career fields. AFOQT Form T version 1 (Tv1) has 10 cognitive subtests, nine of which are scored, including markers of verbal, quantitative, spatial, and perceptual speed abilities, along with Aviation Information (AI) and Instrument Comprehension (IC). The development and maintenance of the AFOQT has been informed by substantial validation research on the test and by the broader literature (e.g., Schmidt & Hunter, 1998) documenting the predictive validity of cognitive abilities for training and job performance outcomes. The scored subtests are combined in multiple ways to produce composite scores for officer commissioning and for classification into rated career fields, including manned and remotely piloted aircraft (RPA) pilot, air battle manager (ABM), and combat systems operator (CSO). Form Tv1 includes one unscored cognitive ability subtest, Physical Science (PS), and two unscored noncognitive subtests, Situational Judgment Test (SJT), and Self-Description Inventory (SDI-O). PS was in scope for this project, but the SJT and SDI-O were not.

Form Tv1 was recently evaluated and found to be psychometrically adequate, but areas of improvement were identified (Kantrowitz et al., 2022; Walsh, Brady et al., 2022; Walsh, Woolley et al., 2022). The mandate for the development of Form T Version 2 (Tv2) was to reduce mean score subgroup differences (SGDs) for gender, racial, and ethnic minority applicants while maintaining psychometric quality (i.e., discriminability, reliability, construct validity, criterion-related validity). The goals for Form Tv2 were to (1) refresh a portion of each subtest (or the entire subtest, in the case of Reading Comprehension) to extend the shelf life of the AFOQT until Form U is deployed and (2) maintain the psychometric quality of the assessments while minimizing SGDs.

New items were developed and trialed alongside some anchor items from Form Tv1. Use of anchor items was designed to facilitate comparison of the Tv1 items to the new items. Data from 70,000+ responses for Form Tv1 were available to compare the psychometric properties of these well-researched items and the new items. The data collection strategy on the new items included research participants from crowdsourcing services and USAF officer candidates, with specific AFOQT subtests targeted to each population. The item analysis strategy used Item Response Theory (IRT) so that Form Tv1 and the new items could be evaluated and calibrated on a common metric. The form assembly methodology (to determine which items to retain for each subtest)

consisted of the generation of multiple combinations of items and evaluation of which combination best met the desired criteria, namely maintenance of score reliability and minimization of mean score SGDs. Recommendations for final subtests were presented for review; upon approval, the assessment content was used to create web-based subtests for Form Tv2. Scores on Form Tv2 were equated to Form Tv1 *t*-scores. Alternative composite scoring formulas (for Pilot, CSO, and ABM) were evaluated. These new alternative composites substantively differed from Form Tv1 and thus required a new approach because the scores would not be considered equivalent or concordant between Form Tv1 and Form Tv2. We recommended a dual-composite approach with super scoring, in which the legacy Form Tv1 and new Form Tv2 composites would be considered in tandem, with the highest scores used as scores of record for selection and classification decision making.

2.0 INTRODUCTION

The research and evaluation of the AFOQT is a model for other testing programs. Maintaining the AFOQT requires substantial effort, analysis, and ongoing research to ensure its utility to the USAF. This technical report describes the development of revised cognitive subtests to support a new version of the AFOQT, Form Tv2. The two non-cognitive assessments (SJT, SDI-O) that are included in the AFOQT Form Tv1 are not in scope for the current effort. There are ten cognitive subtests, nine of which are scored and used in operational composites (PS is not scored). Table 1 provides brief descriptions of each of the AFOQT Form Tv1 cognitive subtests.

Table 1. AFOQT Form Tv1 Subtest Descriptions

Subtest	Description
Verbal Analogies (VA)	Measures the ability to reason and see relationships among words. The examinee must choose the option that best completes the analogy developed at the beginning of each statement.
Arithmetic Reasoning (AR)	Measures the ability to use arithmetic to solve problems. Each problem is followed by five possible answers. The examinee must decide which one of five choices is correct.
Word Knowledge (WK)	Measures knowledge of words and their meanings. 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 and principles. 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) (unscored)	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 block 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.

The AFOQT Form Tv2 development was designed to accomplish several objectives. They were related to the dual goals of refreshing the test content to enable the ongoing use of AFOQT scores and minimizing mean score SGDs for minority candidates. The objectives documented in this report include the following:

Objective 1. Replacement of Problematic AFOQT Form Tv1 Items for Current Cognitive Subtests.

The first objective focused on partially refreshing the AFOQT cognitive subtests to increase the longevity of Form Tv1 through the development of new items for multiple subtests that were designed to reduce known existing SGDs. Note that one test (RC) was fully re-developed rather than partially refreshed to shorten the length of the reading passage stimuli.

Objective 2: Data Collection, Item Analysis, and Scale Development.

The second objective was related to the design, coordination, and execution of a data collection strategy to obtain data on the newly developed items from Objective 1. The strategy included online (web-based) and in-person (web-based and paper-based) data collection using a range of sources, including crowdsourcing research participation and samples of USAF officer candidates. The data collection strategy allowed for sufficient sample size and sample characteristics to enhance generalizability and a range of psychometric item analyses to develop the subtests for Form Tv2. The scale development focused on identification of existing Form Tv1 items and newly developed and calibrated Form Tv2 items that best achieved the goals of maintaining precision (reliability) while reducing mean score SGDs.

Objective 3: Equivalency/Equating Study.

The third objective focused on equating the recommended Form Tv2 subtests with Form Tv1 subtests, given the desire to maintain the meaning of scores from Form Tv1 to Form Tv2. Equating focused on standardized scores (*t*-scores). This way, scores were made comparable across forms and if used, the same cut score could be used across forms. Conversion formulas and tables were provided to facilitate this comparison.

This objective also focused on the development of parallel forms, designed to equate AFOQT Tv1 Form T1 with Tv2 Form A and Tv1 Form T2 with Tv2 Form B. Separate equating procedures were developed.

Objective 4: Develop Two Forms of AFOQT Form Tv2 and Support New Composite Scoring.

The fourth objective focused on finalizing AFOQT Form Tv2 for web-based delivery. Final assessment content and scoring were submitted for delivery on web-based platforms.

This objective also involved the review, revision, and implementation of composite scoring for several existing composites, including the Verbal and Quantitative officer commissioning composites, and existing and newly developed aircrew classification composites for Pilot, CSO, and ABM training. As part of this, we evaluated the potential for an alternative set of composites for Pilot, CSO, and ABM training that incorporates selected facets of the SDI-O. The Air Force Personnel Center Strategic Research and Analysis Branch (AFPC/DSYX) sought a set of alternative composites that could be equated to Form Tv1 composites to maintain score meaning over time/across forms. We evaluated the concordance of the proposed alternative composites with the Form Tv1 composites. We made recommendations for supporting two sets of composites with super scoring to enable candidates to qualify under either composite approach.

3.0 METHOD, ASSUMPTION AND PROCEDURES

In this section, we describe the method used to replace a subset of items within the Form Tv1 subtests, the data collection strategy and process, and analysis methods used to recommend final Form Tv2 subtests.

3.1 Objective 1: Replacement of Problematic AFOQT Form Tv1 Items for Current Subtests

AFOQT Form Tv1 underwent a rigorous evaluation (Kantrowitz et al., 2022; Walsh, Brady et al., 2022a; Walsh, Woolley et al., 2022). The psychometric performance of Form Tv1 items and subtests was reviewed to create an item development roadmap for Form Tv2. This roadmap was translated into a blueprint for item writing to target particular content areas and specific difficulty ranges.

We developed more items than were needed to ensure that the items flagged in the Form Tv1 review were revised. After an iterative item writing and review process, we developed 30 VA, AR, TR, and MK items; 95 RC items; 26 PS items; 43 WK items; 25 IC items; 36 BC items; and 32 AI items. The final item bank contained 377 items across the 10 cognitive subtests.

The new items were developed to target a range of item difficulty and content coverage similar to that encompassed by Form Tv1 subtests. Items included an item stem with one correct answer and four or five response options per item, dependent on Form Tv1 subtest format.

Delivery of items was done in a two-step process to allow the USAF time to review and approve the items. The following item development and review methodology was used to ensure high quality of the item pool:

- Item writers were trained and instructed on the USAF item taxonomies, required item formatting, and sources for generating item content.
- Item writers drafted items in waves beginning with a five-item sample.
- Item reviewers reviewed the item samples to provide feedback and corrections. Items were reviewed by two reviewers (a peer review and a senior team member review). Item writers revised their items per the reviewers' feedback. Item reviewers reviewed the items again to ensure the revisions addressed all feedback. Item reviewers provided additional feedback as needed. Item writers addressed additional rounds of feedback as needed through item revisions.
- All items were reviewed as a set a final time to ensure the draft items fully spanned the intended range of difficulty, were contextualized, and face valid to the military (where appropriate), met the existing content taxonomies, included one and only one correct response, were expected to reduce SGDs and enhance validity, and were sufficiently different from items that appeared on Form Tv1.
- We provided the five-item sample to the AFPC/DSYX prior to continued item writing.

The USAF returned feedback to Personnel Decisions Research Institutes on the samples within 7 business days. We addressed any feedback provided by the AFPC/DSYX and then resubmitted the final items to their approval. Once approved, we continued with additional item writing. We submitted items for AFPC/DSYX review in batches as identified at the outset of the project. We

prepared a deliverable of all items in a phased manner for trialing during data collection in Objective 2.

3.2 Objective 2: Data Collection, Item Analysis, and Scale Development

We designed, coordinated, and executed a data collection strategy that used the Questionmark OnDemand for Government (Questionmark for short) assessment platform. The strategy included options for web-based and paper-based data collection to expedite the data collection and fulfill technical objectives. While it was originally anticipated that the primary data collection mechanism would be to seed the newly developed items into existing AFOQT Form Tv1 operational test administration, alternative strategies were developed that included data collection with three populations: (1) Prolific (a service) (2) Officer Training School (OTS); and (3) Squadron Officer School (SOS). This strategy was used to ensure sufficient sample sizes, sample characteristics, and efficiencies were obtained to enhance generalizability and to facilitate a range of analyses to be performed, including:

- Classical Test Theory (CTT) analysis, such as descriptives, difficulty, discriminability, and mean score SGDs (see Appendix A);
- Item Response Theory (IRT) analysis, comparing newly developed items on the same metric with the existing Form Tv1 items, which was critical given the goal to partially refresh the subtests;
- Analysis to equate Form Tv2 to Form Tv1;
- Form assembly analysis to determine the optimal subset of items for each subtest that best achieved dual goals of subtest reliability and minimizing SGDs for gender, race, and ethnicity.

3.2.1 Identification of Anchor Items for Inclusion in Data Collection

It was determined that a subset of items from Form Tv1 would serve as anchor items to link item-level psychometric properties from previous datasets with the new datasets. Data on both parallel forms ($N_{T1} = 36,613$ and $N_{T2} = 34,254$) were analyzed to identify items that fit target ability distributions to reflect the expectation that the pilot samples were likely to be more variable and contain more low scoring examinees compared to officer candidates who completed Form Tv1. Item difficulty estimates (IRT b -parameters) were calculated in BILOG-MG for items from Forms T1 and T2. For those items in common to both T1 and T2, the b -parameters were stable and highly correlated, as expected. The following information was also produced: Target item information functions (IIF) based on the expected ability distribution, initial IIF based on items eligible to serve as anchor items, and final IIF based on recommended sets of anchor items. The IIF were reviewed to ensure that anchor items represented a range of item difficulty that would be suitable for measuring USAF officer candidates and Prolific study participants. Overall, 10 items were selected per subtest to serve as anchor items in the data collection sessions.

3.2.2 Development of a Data Collection Strategy

Several goals guided the data collection strategy to try out the new items, namely efficiency and representativeness of the samples to the USAF officer candidates. AFOQT Form Tv2 was designed to serve as an interim refresh prior to the next full revision of AFOQT Form U. As such, the USAF

was open to strategies that enabled gathering data rapidly, considering that scientific rigor and sample representativeness were built into the process.

Prolific Data Collection Strategy

The first data collection strategy was to crowdsource participants. To achieve this, we used the Prolific crowdsourcing research platform. Prolific is a web-based research platform that provides recruitment and management of participants for web-based research. Prolific has been shown to result in high quality and more usable data, compared to other crowdsourcing platforms (Peer et al., 2022). Prolific screens prospective research participants and requires participants to register with relevant background information, which enables targeted recruitment strategies, which was important for piloting the AFOQT items. Prolific enables researchers to use virtually any study software to integrate with Prolific studies by embedding a link to the study. Researchers specify sample characteristics and Prolific targets studies to participants who meet the criteria. Once participants complete a study, their data can be easily reviewed and approved to issue compensation.

Prolific provides strong recommendations for participant compensation. To achieve quality data in a reasonable timeline, Prolific provides recommended compensation rates and also suggested that data quality is maximized for studies lasting not longer than approximately 30 minutes. As such, we incorporated this guidance into our strategy. The research forms we developed were designed to take approximately 30 minutes and compensation was set per the Prolific recommendations. We reviewed historical Form Tv1 data to determine completion times for various subtests, to inform which subtest(s) to include in a given research study.

We identified sample characteristics that would maximize generalizability of the tryout research findings with the USAF officer candidate population. As such, we implemented several participant requirements into the Prolific data collection strategy. These included the following: (1) Must be located in the U.S.; (2) minimum age of 18, maximum age of 30; (3) nationality: US; and (4) fluent in English.

We aimed to collect data from 600 participants per subtest. This number was targeted to fulfill sample size requirements for IRT analysis, typically set at 500 participants to achieve accurate parameter estimates, recognizing that data cleaning would result in some amount of data loss.

Next, we considered which of the AFOQT subtests would be amenable to this data collection strategy. We determined that broad ability tests (marker tests of verbal, quantitative, and spatial abilities), a speeded test (IC), and general knowledge tests (PS, AI) were amenable to data collection on Prolific. These tests do not assume military knowledge and do not have unique research needs. As such, the following subtests were targeted to the Prolific data collection strategy: VA, AR, WK, MK, PS, IC, AI, and BC.

We noted that Prolific participants may perform differently than officer candidates on the AI subtest. While this is a general test of aviation knowledge, we recognized that interests and motivation of officer candidates may differ from those of the general population given the relevance of the knowledge domain to Air Force officers. Performance on PS, in contrast, was not expected to differ by sample as this knowledge domain is relevant to a wide range of roles. We

hypothesized that officer candidates may have a higher level of declarative knowledge of aviation as they are pursuing a career in the USAF. Due to this, we also aimed to collect data on this test through other means to compare results from different sources and determine which was the most representative of officer candidates.

USAF Data Collection Strategy

Some tests were anticipated to take longer than the Prolific study length recommendations, had unique research needs, and/or required administration of paper-based and web-based test versions to perform subtest equating.

Given these needs, we worked with AFPC/DSYX to identify relevant and available research samples. The first sample consisted of enlisted personnel attending Basic Military Training (BMT) at the Applied Performance Assessment and Testing (APAT) center at Lackland Air Force Base (AFB). It was noted, however, that BMTs were likely to differ from officer candidates and that internet connectivity was not yet available at the APAT, but a request was made to procure and enable the connectivity. Despite the concern about sample characteristics and internet connectivity, this was viewed as a provisional strategy for collecting data in person. Data collection kicked off with several participant samples, but the turnout was very low. Also, it was determined that the timeline for procuring and installing internet connectivity was likely to be lengthy. As such, we worked with AFPC/DSYX to obtain access to other research samples.

AFPC/DSYX worked with USAF leadership and determined that classes attending OTS and SOS that were amenable to providing samples for this research. The classes were asked to attend a research session lasting approximately 3 hours. The classes were randomly assigned to administration of paper-based and web-based test forms, which was used to achieve a random equivalent group design. We aimed to collect data from 300 participants per each condition. The data collection sessions for Form A were conducted in June 2022 and for Form B in November 2022.

The subtests targeted for USAF data collection included the following, along with rationale for inclusion:

- **RC:** This test was intended to be fully rather than partially refreshed; as such, more items needed to be tried out compared to the other tests resulting in a long assessment that was not conducive with Prolific study length recommendations to achieve high quality data.
- **TR:** to establish equivalence for this speeded test, we needed to administer the test in paper- and web-based forms. Speeded tests are known for being challenging to equate, due to differences in speed associated with different response formats (bubbling a Scantron sheet for paper-based tests, clicking an option for web-based assessment). As such, random administration of the paper-based and web-based assessments was designed to provide the data to establish a rate scoring approach, in which the proportion of correct responses corrected for guessing is divided by the mean screen presentation time for the items (Segall et al., 1997). In addition to the test equating requirement, new items were developed and included in the study to evaluate the performance of the tests.
- **IC:** This subtest was targeted for inclusion in the data collection with USAF samples after the Prolific study results from Form A showed a large difference in test performance with

Form Tv1 data. The anchor items were compared and large differences in difficulty were noted. The presentation of the items also differed in the web-based versus paper-based formats. Following the Prolific data collection for Form A, the item format was updated to more closely mimic the original paper-based version to achieve comparability with the Form Tv1 paper-based results.

- **BC:** As noted above, BC was included in the Prolific data collection strategy in order to collect data and evaluate the new items. In addition, the research plan for BC also involved equating web-based to paper-based stimuli (for the same items), in which (the new items used highlighted (colored) block towers and the former version used numbered block towers.
- **AI:** as noted above, we hypothesized that individuals aspiring to a career in the USAF may have higher levels of aviation knowledge, compared to the general public. Given the job knowledge nature of this assessment, we also included this subtest in data collection with the USAF samples.

3.2.3 Preparing Test Materials

We worked with a government contractor, 1CyberForce, who was responsible for developing and administering the web based AFOQT, to create the web-based assessment forms on Questionmark. This involved providing them with the new and anchor item content along with subtest level specifications (e.g., test length, navigation features) and ensuring that the test forms were built to specifications and functioned as intended. In the process of setting up and finalizing the pilot assessments, 1CyberForce determined that one of the intended data collection strategies (Prolific) required review for security considerations. As noted earlier, we anticipated providing a link to research participants to complete the assessments. This required an open assessment link on the Questionmark platform. The use of open assessment links was flagged by 1CyberForce for potential security issues. In the meantime, to mitigate timeline risk, an alternative method was pilot tested, which involved a two-stage process in which Prolific participants (1) registered for the study and provided their Prolific e-mail addresses and (2) received the links to the assessments. This alternative method did not yield sufficient participation in Step 2 so was discarded as a viable data collection strategy.

The open assessment approach was eventually approved but caused a significant delay in the overall timeline. In addition to the open link, another security feature offered by Questionmark is a browser lockdown. This requires participants to download a secure browser before beginning the assessments. Once downloaded, the assessments are launched in a manner that does not permit the use of other software programs or websites. We elected to use this feature to minimize the opportunity for participants to use resources like calculators and websites to search for answers.

In the design of the materials for the Prolific samples, we also created and embedded several data quality filter items that were designed to help diagnose whether participants were using other resources, paying attention, and putting forth high effort. These items varied by subtest and were designed to match the style of the items in that subtest to avoid detection. For several subtests, we included easy items that anyone who understood the task should have gotten correct. Since AR and MK do not currently permit the use of calculators, we instructed research participants to avoid the use of calculators. To detect the use of calculators, we included one question in each subtest

that contained a correct answer if solved as intended and one answer that was almost correct and could have only been solved with a calculator. These questions were intended to be used as data quality filters.

A short post-test survey was developed that was designed to measure self-reported effort on the study, the use of external resources, and whether the data should be used in the study.

Several demographic items were included, such as age, gender, ethnicity, race, socioeconomic status, education, highest degree attained, recruitment source, English language fluency, and zip code associated with home address.

Finally, USAF-provided language regarding informed consent was also included in the test materials.

Paper booklets were created for the in-person/paper-based testing condition. The booklets were created to resemble the ones used in Form Tv1 with the same cover page (albeit with a different booklet identification number). The web-based and paper-based test forms were created and thoroughly reviewed for quality assurance.

3.2.4 Data Collection: AFOQT Form Tv2 A and B

Prolific Studies

We developed three studies (per Forms A and B) intended for Prolific data collection that fit the intended 30-minute duration for test administration. The three studies and subtests were as follows:

- AFOQT Form Tv2 (A):
 - o Prolific 1 (A): AR (26 items, estimated duration = 30 minutes)
 - o Prolific 2 (A): MK (26 items), PS (24 items) (estimated duration = 30 minutes)
 - o Prolific 3 (A): VA (26 items), WK (40 items), IC (26 items), BC (28 items), AI (24 items) (estimated duration = 40 minutes)
- AFOQT Form Tv2 (B):
 - o Prolific 1 (A): AR (26 items, estimated duration = 30 minutes)
 - o Prolific 2 (A): MK (26 items), PS (22 items) (estimated duration = 30 minutes)
 - o Prolific 3 (A): VA (26 items), WK (25 items), BC (28 items), AI (28 items) (estimated duration = 30 minutes).

For Tv2 (A), we conducted an initial tryout with 50 participants to ensure the process was seamless and to solicit reactions from participants. Following this, we increased the total intended sample size to 600 for each of the three studies. We tracked the usage and participant questions throughout the process. Data were generally collected in approximately a two-week time span over March and April 2022. For Tv2 (B), data were collected between June and July 2022. Note that for Tv2 (B), we did not include IC as large differences were found on the anchor items when comparing USAF officer candidates (based on Form Tv1 data) and Prolific participants. As described later in this report, these data were not used in analysis; as such, a decision was made to exclude this subtest in the Prolific Form B data collection.

The web-based and paper-based assessment battery was expected to take approximately 3 hours to complete. Participants were randomly assigned to the web-based or paper-based condition. For Tv2 (A) data were collected in June 2022. For Tv2 (B) data were collected in November 2022. Target sample sizes per subtest for each version were 600 participants (300 for paper-based data collection and 300 for web-based data collection). Table 2 shows the study name, subtest name, date of data collection, target sample size (N) and final sample size (prior to cleaning).

Table 2. Details of Data Collection for AFOQT Form Tv2 (A and B)

Study	Subtest(s) Included	Date of Data Collection	Target N
Prolific 1 – A	AR	March 2022	600
Prolific 2 – A	MK PS	April 2022	600
Prolific 3 – A	VA WK IC BC AI	April 2022	600
OTS – A and SOS - A	RC TR AI IC BC	June 2022	300 (web) 300 (paper)
Prolific 1 – B	AR	June 2022	600
Prolific 2 - B	MK PS	July 2022	600
Prolific 3 – B	VA WK BC AI	July 2022	600
OTS – B and SOS - B	RC TR AI IC BC	November 2022	300 (web) 300 (paper)

3.2.5 Item Analysis Procedures

After data cleaning, (reported in Section 3), item analyses were conducted using two sets of procedures: (1) CTT analyses and (2) IRT analyses. Given the goal to partially refresh the AFOQT for Form Tv2 through the use of a mix of Form Tv1 and newly developed items, and that IRT allows for evaluation of items on a common metric, IRT was the primary means for item analysis. Interested readers can refer to the Appendix and Attachment 1 “CTT Parameters.xlsx” for more information on the CTT item-level analysis results.

For the IRT analysis, a series of steps was taken to evaluate item-level performance. The first step was to assess model-data fit. The assessment of model-data fit is used to determine if a

measurement model fits the data well enough in an absolute sense to use the study participant parameter estimates. When a model is not appropriate or does not fit the data, the use of the derived item parameter may be compromised (Stone & Zhang, 2003). BILOG-MG (Zimowski et al., 2003) requires a specific preparation of the item response data. The scored assessment responses were coded as correct answer = 1; incorrect answer or omissions = 0; and not administered = 9. Data from the experimental OTS, SOS, and Prolific samples were combined with historical AFOQT data. The experimental and historical data were coded so that the historical data could be used as the reference group for BILOG to create parameters. This was done so BILOG would create parameters for the experimental items that were on the same scale as previous (historical) AFOQT assessments. The recoded data were written to an ASCII file for analysis using BILOG-MG 3.0 with default parameters except that additional estimation cycles were allowed, threshold priors were requested, convergence criterion for expectation maximization (EM) and Newton iterations was modified on a case-by-case basis to ensure fit (up to 1.0 but typically < .01), and a not-administered key was supplied.

In addition to parameterizing the items, Differential Item Functioning (DIF) analyses were conducted on all items. DIF is used to detect interactions between item difficulties and protected minority subgroups (Thissen et al., 1993). This analysis allows investigation into whether members of subgroups with the same latent ability have the same probability of a correct response as those in the majority group. Since DIF alters the meaning of a test score between one group and another, it is desirable to remove DIF, if possible. While DIF can also be analyzed using CTT statistics, such as the Mantel-Haenszel (MH) statistic, IRT estimations of DIF tend to be more sensitive, especially in shorter tests, because IRT better defines the latent construct measured by the test. Therefore, CTT MH DIF analyses were not attempted. In BILOG, it is assumed that DIF is restrained to difficulties only and does not apply to item discrimination. Parameters for the separate groups are estimated based on the assumption that the slope parameters are homogeneous across groups.

3.2.6 Form Assembly Goals and Process

The IRT parameters were used to generate many possible test forms consisting of Form Tv1 items and newly developed items. The forms were generated and evaluated to ensure that the set of items reduces SGDs (as compared to Form Tv1) while maintaining adequate test precision.

Diversity Index

As stated previously, the goals for scale development included minimal loss of precision (subtest reliability) and minimization of SGDs. Because development of Form Tv2 was intended to be based on a mixture of Form Tv1 items and newly developed items, and because matched data were not collected or available on all potential items to calculate mean score SGDs, a procedure needed to be developed to estimate the potential for enhanced diversity. We based subgroup definitions on U.S. Office of Management and Budget (OMB) standards, which refer to mutually exclusive racial and ethnic (non-overlapping) categories. It was noted that the same definitions are used by the Armed Services Vocational Aptitude Battery (ASVAB) and is used by much of the Department of Defense (DoD) (e.g., Defense Manpower Data Center, DMDC). The U.S. Census Bureau conceptualized a diversity index (DI) and was reported for the 2020 Census; the DI is equal to the probability that two people chosen at random will be from different race and ethnic groups. In our

approach, the DIs for unselected (applicant) and selected (qualified) groups of different test forms were calculated and compared. (The qualified group was defined as the group scoring at or above the mean on the current or revised test form; the mean may represent a conservative estimate of qualification). Subgroup calculations were conducted and reported for gender and race/ethnicity).

Form Assembly

The form assembly process was determined based on estimated IRT parameters for new and existing items. First, the latent ability distributions were estimated based on Form Tv1 data from officer candidate samples for each subgroup defined above by racial, ethnic, and gender groups. The latent distributions represent the true distributions of latent ability and are defined (and scaled) independently of the test-items and the test-scores they produce. These distributions represent the current state of subgroup ability differences on the subtests and serve as the reference point for alternative forms that use a combination of Form Tv1 items and newly developed items. Next, subgroup performance on alternative forms was estimated using IRT. Many potential alternative forms (approximately 100,000) were created to evaluate which form(s) most substantially increased the diversity index while maintaining adequate test precision.

3.3 Objective 3: Equivalency/Equating Study

As noted earlier, scores on Form Tv2 were intended to be comparable to scores on Form Tv1. As such, a method to equate scores on the new and old versions was necessary. Form Tv1 scores were reported in terms of t -scores (with a mean [M] of 50 and standard deviation [SD] of 10). To equate scores from Form Tv2 to Form Tv1 (which were previously equated to prior AFOQT versions), we used the IRT observed score method. In this method, IRT parameters were used to generate number right distributions for T1 and Tv2 (A) and T2 and Tv2 (B), respectively, which were used to transform number right scores from Tv2 (A) to T1 and Tv2 (B) to T2, respectively. The equated t -scores were then amenable to percentile conversions and inclusion in composite score calculations.

3.4 Objective 4: Develop Two Forms of AFOQT Form Tv2 and Support New Composite Scoring

The method for developing final test forms involved several logistical steps, including implementing the recommended subtests on the respective web-based platform (Questionmark in the case of Tv2 (A) and Pearson Vue in the case of Tv2 (B)).

To facilitate the implementation of the new subtests, updates to scoring protocol were required, specifically to use equated t -scores for Form Tv2 subtests in composite scoring formulas for the Verbal, Quantitative, Academic Aptitude, Pilot, CSO, and ABM composites. Formulas and tables that replicated the T1 and T2 composite calculations were constructed for use with Tv2 (A) and Tv2 (B). These formulas use t -scores from the latter (new) forms in place of t -scores from the former forms (T1 and T2).

It was also determined that the USAF desired to implement updated composite scoring for Pilot, CSO, and ABM composites that included selected facets from the SDI-O. See Woolley et al. (2023) for more details on the alternative composites. Alternative composites were determined

based on improved validity outcomes compared to Form Tv1 composites. However, it was also desired to preserve the ability to equate composite scores between previous composite scores and Form Tv2 scores. Given the dissimilarity between legacy composites (which did not include SDI-O facets) and new composites (which included SDI-O facets), it was not assumed or expected that the composites could be considered equatable. Rather, evaluation of concordance was proposed. Concordance (rather than equating) was potentially suitable, given its intended use when tests measure dissimilar things but may be considered corresponding/concordant rather than interchangeable. As such, a method was developed to determine the extent to which old and new composites were comparable and could be linked.

At the outset, we recognized several competing technical and practical goals for AFOQT Tv2. These included technical goals which related to (1) revising the cognitive subtests to reduce mean score SGDs and refreshing the content and (2) implementing SDI-O facets in scored composites to enhance validity. AFPC/DSYX also had several practical goals (1) equating Form Tv2 to Form Tv1 scoring to maintain use of current cut-scores and eligibility of previous candidates scores and (2) supporting Form Tv1 paper-based testing and Form Tv2 web-based testing simultaneously. It was noted that the alternative composites would not allow for equating to Form Tv1, as a different test (namely, the SDI-O) was intended to be a scored component, whereas it was not in the Form Tv1 composites. That is, scores would not be interchangeable but could potentially be considered corresponding/concordant. Data needed to be evaluated to determine whether scores correspond.

The method for feasibility of score linking was developed to consider the following: (1) construct and test specification similarity, (2) similarity of inter-correlations, reliability, and validity, (3) subpopulation invariance, (4) evaluation of linking methods (linear, equipercentile), (5) the extent to which faking (which may become more prevalent after disclosure to examinees that SDI-O is a scored rather than experimental component of AFOQT) may alter percentile rank.

4.0 RESULTS AND DISCUSSIONS

4.1 Objective 1: Replacement of Problematic AFOQT Form Tv1 Items for Current Subtests

The number of items developed and tried out for each subtest is shown in Table 3.

Table 3. Counts for Items Developed and Pilot Tested

Subtest	Count of Tryout Items	
	Tv2 A	Tv2 B
VA	15	15
AR	15	15
WK	29	14
MK	15	15
RC	45	50
PS	14	12
TR	15	15
IC	15	10
BC	18	18
AI	14	18

4.2. Objective 2: Data Collection, Item Analysis, and Scale Development

4.2.1 Data Cleaning

Prior to conducting data analyses, it is considered a best practice to clean the data to ensure the analyses are based on quality data.

First, we merged the raw data generated by Questionmark (web-based) or a Scantron reader (paper-based) in various ways in Excel. In particular, each subtest was merged in such a way to include the subtest items, as well as a copy of the demographics and post-test surveys. Each resulting Excel dataset was further scrubbed to remove extraneous meta-data and was eventually imported into R.

Next, in R, we checked the accuracy of the web-based generated scores against the scoring keys. The raw web-based item-level data were rescored (using the scoring keys) and aggregated into a subtest score. This newly calculated subtest scores were then correlated with the “total score” column provided by Questionmark. Correlation coefficients of a 1.0 represented a match between the two scoring methods. Correlations below 1.0 were closely examined and responses favoring the official scoring key were retained. This occurrence was quite rare, with a vast majority of the subtests being scored correctly both from Questionmark and manually. Note that the raw paper-based item-level data were also manually rescored in R (using the same scoring keys as for the web-based data) and aggregated into a subtest score. However, no subtest-level score was provided in the original dataset so checking score accuracy was not possible.

Then, we examined unit missingness (operationalized as the absence of any responses for the entire subtest) and unit longstring (operationalized as consecutive endorsement of the same response option for the entire subtest) to identify participants who may not have paid attention or demonstrated high effort. The raw web-based and paper-based item-level data were recoded into A, B, C, D, or E when a 5-point response scale was used or into A, B, C, or D when a 4-point response scale was used (i.e., IC).

Finally, we examined chance responding to identify participants who may not have paid attention. Chance responding is operationalized as a minimum subtest-level score, at or below which a participant can endorse items correctly purely by chance. The chance responding threshold was calculated as a total number of items divided by number of response options. For example, a subtest with 30 items and 5 response options would receive a minimum score based on chance responding of 6. Those who scored less than or equal to 6 would be considered careless cases and would be removed listwise. Chance responding was examined separately for each subtest.

Applying the aforementioned data cleaning techniques resulted in a reduction of the original sample sizes. The lowest reduction was observed for Prolific WK (2%) and the highest reduction was observed for OTS/SOS TR (25%). The average sample size reductions across subtests and different samples was 11%.

Final sample sizes representing cleaned data are shown in Table 4.

Table 4. Sample Sizes after Data Cleaning

Subtest	Tv2 (A)	Tv2 (B)
VA	608	602
AR	582	600
WK	605	600
MK	516	607
RC	620	616
PS	578	603
TR	637	610
IC	285	620
BC	589	591
AI	1,169	1,196

4.2.2. IRT Analysis

For all assessments subjected to IRT analysis (i.e., AI, AR, BC, IC, MK, PS, RC, VA, and WK), the EM cycles converged in 75 cycles or less and the Newton-Raphson cycles converged in 30 or less additional cycles. Item parameter estimates for items in each assessment are included in Attachment 2 “IRT Parameters.xlsx.”

To assess DIF, BILOG provides an overall marginal likelihood ratio test. To perform this test, the data is analyzed in a single group as if they came from the same population, and the marginal maximum log likelihood of the item parameters is recorded. Then, the data are analyzed in separate groups using the DIF model and the final log likelihood is recorded. The difference in these log likelihoods is distributed in large samples as χ^2 with $(n - 1)(m - 1)$ degrees of freedom, where n is the number of items and m is the number of groups. When this χ^2 is significant, there is evidence that DIF is present. Based on these analyses, no items showed evidence of DIF.

4.2.3 Form Assembly Results

Initial analysis focused on existing latent ability distributions (by race, ethnicity, and gender) based on officer candidate data from AFOQT Form T1 and T2. Results were expressed in terms of M and SD values in theta scoring terms, in which the expected M is 0 and SD is 1. Negative values indicate that a given subgroup scored lower than average, and positive values indicate that a given subgroup scored higher than average. The latent ability M for the White Non-Hispanic groups was higher than Non-White groups across all subtests. That is, scores for White Non-Hispanics were higher than for other groups. Given these results, to more accurately assess the Non-White subgroups, assessments for Form Tv2 required multiple items that fall within the latent ability distributions of the target, Non-White groups to drive reductions in SGDs compared to Form Tv1. The same observation was noted for gender differences in which the latent ability distributions showed higher mean results for males compared to females. See Table 5.

Table 5. Latent Ability Distribution Average Values based on Form Tv1 and Tv2 Analysis

Subtest	Hispanic/Latino		White Non-Hispanic/Latino		Black or African American Non-Hispanic		American Indian and Alaskan Native Non-Hispanic		Asian Non-Hispanic		Native Hawaiian and other Pacific Islander Non-Hispanic		Some Other Race alone Non-Hispanic		Multiracial Non-Hispanic		Male		Female	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
VA	-.34	1.04	.29	.94	-.86	1.08	-.21	1.07	-.25	1.05	-.58	1.04	.21	.98	-.02	.91	.07	.95	-.24	1.08
AR	-.34	1.01	.22	.93	-.90	1.00	-.34	1.23	.22	1.09	-.59	.93	-.03	1.04	-.11	.92	.15	1.00	-.52	1.05
WK	-.31	1.00	.25	.95	-.70	1.01	-.26	1.07	-.25	1.12	-.44	.86	.23	1.09	-.02	.93	.07	1.03	-.26	.99
MK	-.27	.98	.15	.98	-.70	.95	-.25	1.03	.51	1.02	-.30	.95	-.05	1.06	-.08	.90	.09	1.04	-.31	.99
RC	-.25	1.09	.26	.93	-.79	.94	-.34	1.10	-.24	1.01	-.58	.99	.10	1.00	-.02	.97	.12	1.01	-.44	.92
PS	-.34	1.03	.21	.89	-.97	1.03	-.29	1.17	.03	.96	-.71	1.05	.11	.88	-.16	.98	.17	1.00	-.53	.93
IC	-.24	.97	.20	.97	-.97	.95	-.03	1.30	-.28	.95	-.56	1.02	.06	.94	-.11	.93	.26	.91	-.70	.85
BC	-.22	1.08	.16	.99	-.91	1.09	-.11	1.19	-.02	1.03	-.31	.96	.01	.97	-.09	1.02	.20	.88	-.51	1.13
AI	-.35	1.07	.22	.81	-1.03	1.21	-.05	.93	-.31	.90	-.36	.89	.11	1.10	-.14	.97	.18	.83	-.75	1.05

Note. *M* = Mean; *SD* = Standard Deviation

The form assembly routine generated large numbers of possible solutions for each subtest. Results were reviewed for solutions that maximized reliability and minimized subgroup differences. The form assembly routine was constrained to select forms with no more than a .05 reduction in reliability compared to the reliability estimate for each Form Tv1 subtest. Results were evaluated in terms of improvements in the diversity index, described next.

4.2.4 Diversity Index Results

The diversity indices (DIs) were computed based on AFOQT Form Tv1 data for each subtest and for each form (T1 and T2). The DIs were compared to 2020 census data and applicant group data describing racial, ethnic, and gender diversity. Applicant group diversity was based on all officer candidate demographic data reported on Forms T1 and T2. Form T1 and T2 diversity results were computed by assuming applicant qualification based on the mean of each subtest score, though this is conceptual as cut-scores are not applied at the subtest level. This enabled comparisons with potential Form Tv2 versions to determine which combination of items most substantially improved diversity outcomes. The degree to which Form Tv2 recommended subtests closed the diversity gap between qualified candidates and the applicant group was calculated for each subtest.

Results are shown in Table 6 for aggregate level results and Table 7 for subgroup-specific results. The values in the tables show the extent to which Form Tv2 subtests are expected to close the diversity gap with the applicant pool. For example, based on Form Tv1 data, females comprised 38.5% of the applicant sample, but for T1 AR, only 28.0% of the qualified sample (using the mean as the theoretical threshold for qualification, recognizing that applicants are not selected at the subtest level) was female. So, the existing gap for Form Tv1 is 10.5% (38.5-28.0). Form Tv2 (A) for AR shows that 30.1% of the qualified applicant pool would be female. Compared to Form T, this is an improvement of 2.1% (30.1-28.0). Expressed as the degree to which Form Tv2 closes the gap between the qualified group and applicant group: $(2.1/10.5) * 100 = 20\%$.

Table 6. Diversity Index Results for Race and Gender - Form T1, Form Tv2 (A), Form T2, Form Tv2 (B), and Increase in Diversity for Forms Tv2 (A and B)

Subtest	Category	2020 Census Diversity	Unselected Applicant Group Diversity	Form T1 Diversity ¹	Form Tv2 (A) Diversity ¹	% Change in Qualified Candidate Diversity for Form Tv2(A)	Form T2 Diversity ¹	Form Tv2 (B) Diversity ¹	% Change in Qualified Candidate Diversity for Form Tv2(B)
VA	Race	61.1	60.6	47.6	48.6	7.7	49	50.2	10.3
	Gender	50	38.5	33.9	34.1	4.4	33.1	33.9	14.8
AR	Race	61.1	60.6	50.2	51.5	12.5	48.1	49.5	11.2
	Gender	50	38.5	28	30.1	20	26.8	28.1	11.1
WK	Race	61.1	60.6	48.9	50.3	12	47.6	48.4	6.2
	Gender	50	38.5	33.1	34	16.7	33.9	34.1	4.3
MK	Race	61.1	60.6	53.5	54.3	11.3	52.2	53	9.5
	Gender	50	38.5	32.2	33.1	14.3	31.1	32	12.2
RC	Race	61.1	60.6	48.6	50	11.7	49.2	50.5	11.4
	Gender	50	38.5	28.1	29.9	17.3	28.9	30.5	16.7
PS	Race	61.1	60.6	48.7	49.8	9.2	46.9	48.3	10.2
	Gender	50	38.5	26.9	28.6	14.7	23.7	26	15.5
IC	Race	61.1	60.6	48.3	50.3	16.3	47.9	49.5	12.6
	Gender	50	38.5	19.5	22.6	16.3	18.9	21.4	12.8
BC	Race	61.1	60.6	51.5	52.7	13.2	51.4	53	17.4
	Gender	50	38.5	27.5	28.6	10	27.3	29	15.2
AI	Race	61.1	60.6	45.8	46.5	4.7	45.7	46.2	3.4
	Gender	50	38.5	20.7	21.6	5.1	20.5	21.4	5

Note. ¹ Calculated based on scoring at the mean or higher.

Table 7. Diversity Index Results by Subgroup Race/Ethnicity – Tv2 (A) and (B) and Changes Relative to Form T1 and T2

Subtest	Subgroups	Unselected Applicant Group Diversity	Form T1 Diversity ¹	Form Tv2 (A) Diversity ¹	% Change in Qualified Candidate Diversity for Form Tv2(A)	Form T2 Diversity ¹	Form Tv2 (B) Diversity ¹	% Change in Qualified Candidate Diversity for Form Tv2(B)
VA	Hispanic/Latino	14.60	10.30	10.60	7.00	10.30	10.50	4.70
	White Non-Hispanic/Latino	59.30	70.90	70.00	-7.80	70.80	70.20	-5.20
	Black or African American Non-Hispanic	9.30	3.80	4.10	5.50	3.80	4.00	3.60
	American Indian and Alaskan Native Non-Hispanic	.20	.10	.10	.00	.10	.10	.00
	Asian Non-Hispanic	5.20	4.10	4.10	.00	4.10	4.10	.00
	Native Hawaiian and other Pacific Islander Non-Hispanic	.40	.20	.20	.00	.20	.20	.00
	Some Other Race alone Non-Hispanic	1.40	1.60	1.60	.00	1.60	1.60	.00
	Multiracial Non-Hispanic	9.60	9.10	9.20	20.00	9.10	9.20	20.00
	Male	74.00	78.40	78.20	-4.50	78.30	78.20	-2.30
	Female	26.00	21.60	21.80	4.50	21.70	21.80	2.30
AR	Hispanic/Latino	14.60	10.80	11.00	5.30	10.20	10.70	11.40
	White Non-Hispanic/Latino	59.30	68.80	67.70	-11.60	70.40	69.30	-9.90
	Black or African American Non-Hispanic	9.30	4.20	4.70	9.80	3.10	3.60	8.10
	American Indian and Alaskan Native Non-Hispanic	.20	.20	.10	--	.10	.10	.00
	Asian Non-Hispanic	5.20	5.80	5.50	-50.00	6.10	5.90	-22.20
	Native Hawaiian and other Pacific Islander Non-Hispanic	.40	.20	.30	50.00	.20	.20	.00
	Some Other Race alone Non-Hispanic	1.40	1.30	1.50	200.00	1.40	1.40	--
	Multiracial Non-Hispanic	9.60	8.70	9.20	55.60	8.50	8.80	27.30
	Male	74.00	83.20	81.60	-17.40	84.10	83.10	-9.90
	Female	26.00	16.80	18.40	17.40	15.90	16.90	9.90
WK	Hispanic/Latino	14.60	10.60	11.00	10.00	10.60	11.00	10.00

	White Non-Hispanic/Latino	59.30	69.80	68.70	-10.50	69.80	68.80	-9.50
	Black or African American Non-Hispanic	9.30	4.40	4.80	8.20	4.40	4.80	8.20
	American Indian and Alaskan Native Non-Hispanic	.20	.10	.10	.00	.10	.10	.00
	Asian Non-Hispanic	5.20	4.20	4.20	.00	4.20	4.20	.00
	Native Hawaiian and other Pacific Islander Non-Hispanic	.40	.20	.20	.00	.20	.20	.00
	Some Other Race alone Non-Hispanic	1.40	1.60	1.50	-50.00	1.60	1.60	.00
	Multiracial Non-Hispanic	9.60	9.10	9.30	40.00	9.10	9.30	40.00
	Male	74.00	79.00	78.30	-14.00	79.00	78.40	-12.00
	Female	26.00	21.00	21.70	14.00	21.00	21.60	12.00
MK	Hispanic/Latino	14.60	11.30	12.00	21.20	10.90	11.30	10.80
	White Non-Hispanic/Latino	59.30	66.10	65.30	-11.80	67.20	66.50	-8.90
	Black or African American Non-Hispanic	9.30	4.90	5.30	9.10	3.90	4.30	7.40
	American Indian and Alaskan Native Non-Hispanic	.20	.20	.10	.00	.10	.10	.00
	Asian Non-Hispanic	5.20	7.10	6.80	-15.80	7.60	7.20	-16.70
	Native Hawaiian and other Pacific Islander Non-Hispanic	.40	.20	.30	50.00	.30	.30	.00
	Some Other Race alone Non-Hispanic	1.40	1.40	1.40	.00	1.40	1.40	.00
	Multiracial Non-Hispanic	9.60	8.80	8.80	.00	8.70	8.90	22.20
	Male	74.00	79.80	79.00	-13.80	80.80	80.00	-11.80
	Female	26.00	20.20	21.00	13.80	19.20	20.00	11.80
RC	Hispanic/Latino	14.60	11.40	11.70	9.40	11.50	11.80	9.70
	White Non-Hispanic/Latino	59.30	70.00	68.80	-11.20	69.50	68.40	-10.80
	Black or African American Non-Hispanic	9.30	3.70	4.20	8.90	3.90	4.40	9.30
	American Indian and Alaskan Native Non-Hispanic	.20	.10	.10	.00	.10	.10	.00

	Asian Non-Hispanic	5.20	4.10	4.20	9.10	4.20	4.30	10.00
	Native Hawaiian and other Pacific Islander Non-Hispanic	.40	.20	.20	.00	.20	.20	.00
	Some Other Race alone Non-Hispanic	1.40	1.40	1.50	.00	1.50	1.50	.00
	Multiracial non-Hispanic	9.60	9.10	9.30	40.00	9.20	9.30	25.00
	Male	74.00	83.10	81.70	-15.40	82.50	81.20	-15.30
	Female	26.00	16.90	18.30	15.40	17.50	18.80	15.30
PS	Hispanic/Latino	14.60	11.20	11.60	11.80	10.40	10.80	9.50
	White Non-Hispanic/Latino	59.30	70.00	69.00	-9.30	71.40	70.30	-9.10
	Black or African American Non-Hispanic	9.30	3.50	3.60	1.70	2.90	3.30	6.30
	American Indian and Alaskan Native Non-Hispanic	.20	.20	.10	.00	.10	.10	.00
	Asian non-Hispanic	5.20	5.20	5.60	.00	5.30	5.30	.00
	Native Hawaiian and other Pacific Islander Non-Hispanic	.40	.20	.20	.00	.10	.20	33.30
	Some Other Race alone Non-Hispanic	1.40	1.40	1.40	.00	1.50	1.50	.00
	Multiracial Non-Hispanic	9.60	8.40	8.50	8.30	8.20	8.40	14.30
	Male	74.00	84.00	82.70	-13.00	86.20	84.70	-12.30
	Female	26.00	16.00	17.30	13.00	13.80	15.30	12.30
IC	Hispanic/Latino	14.60	11.90	12.40	18.50	11.80	12.20	14.30
	White Non-Hispanic/Latino	59.30	70.20	68.50	-15.60	70.40	69.20	-10.80
	Black or African American Non-Hispanic	9.30	3.00	3.60	9.50	2.90	3.40	7.80
	American Indian and Alaskan Native Non-Hispanic	.20	.20	.20	.00	.20	.10	.00
	Asian non-Hispanic	5.20	4.10	4.30	18.20	4.10	4.30	18.20
	Native Hawaiian and other Pacific Islander Non-Hispanic	.40	.20	.20	.00	.20	.20	.00
	Some Other Race alone Non-Hispanic	1.40	1.50	1.50	.00	1.50	1.50	.00
	Multiracial Non-Hispanic	9.60	8.90	9.20	42.90	8.90	9.10	28.60

	Male	74.00	89.10	87.00	-13.90	89.40	87.80	-10.40
	Female	26.00	10.90	13.00	13.90	10.60	12.20	10.40
BC	Hispanic/Latino	14.60	12.30	12.60	13.00	12.30	12.70	17.40
	White Non-Hispanic/Latino	59.30	67.60	66.60	-12.10	67.70	66.40	-15.50
	Black or African American Non-Hispanic	9.30	3.90	4.40	9.30	3.80	4.50	12.70
	American Indian and Alaskan Native Non-Hispanic	.20	.20	.10	.00	.20	.20	.00
	Asian Non-Hispanic	5.20	5.20	5.20	.00	5.20	5.20	.00
	Native Hawaiian and other Pacific Islander Non-Hispanic	.40	.30	.30	.00	.30	.30	.00
	Some Other Race alone Non-Hispanic	1.40	1.40	1.40	.00	1.40	1.50	.00
	Multiracial Non-Hispanic	9.60	9.10	9.30	40.00	9.10	9.30	40.00
	Male	74.00	83.50	82.70	-8.40	83.70	82.40	-13.40
	Female	26.00	16.50	17.30	8.40	16.30	17.60	13.40
AI	Hispanic/Latino	14.60	10.50	10.60	2.40	10.50	10.60	2.40
	White Non-Hispanic/Latino	59.30	72.20	71.70	-3.90	72.30	71.90	-3.10
	Black or African American Non-Hispanic	9.30	3.40	3.60	3.40	3.40	3.50	1.70
	American Indian and Alaskan Native Non-Hispanic	.20	.10	.20	100.00	.10	.20	100.00
	Asian Non-Hispanic	5.20	3.60	3.70	6.30	3.50	3.70	11.80
	Native Hawaiian and other Pacific Islander Non-Hispanic	.40	.20	.20	.00	.20	.20	.00
	Some Other Race alone Non-Hispanic	1.40	1.60	1.50	-50.00	1.60	1.50	-50.00
	Multiracial Non-Hispanic	9.60	8.40	8.50	8.30	8.40	8.40	.00
	Male	74.00	88.30	87.70	-4.20	88.40	87.80	-4.20
	Female	26.00	11.70	12.30	4.20	11.60	12.20	4.20

Note. ¹ Calculated based on scoring at the mean or higher

4.2.5 Reliability Estimates

It was expected that reliability would be somewhat, though minimally, compromised as a result of maximizing item information for latent ability distributions for non-White male demographic subgroups. This is due to expected loss in information for the sample as a whole as the majority was composed of individuals who identified as White non-Hispanic male. Internal consistency reliability estimates were evaluated for the large number of alternative solutions generated according to the level of decrements observed. The recommended Form Tv2 subtest reliability estimates are shown in Table 8, along with the current reliability estimates for AFOQT Forms T1 and T2 and the difference between the Form Tv1 and Form Tv2 values. In all cases, the resulting change in reliability was minimal (to the hundredth of a decimal).

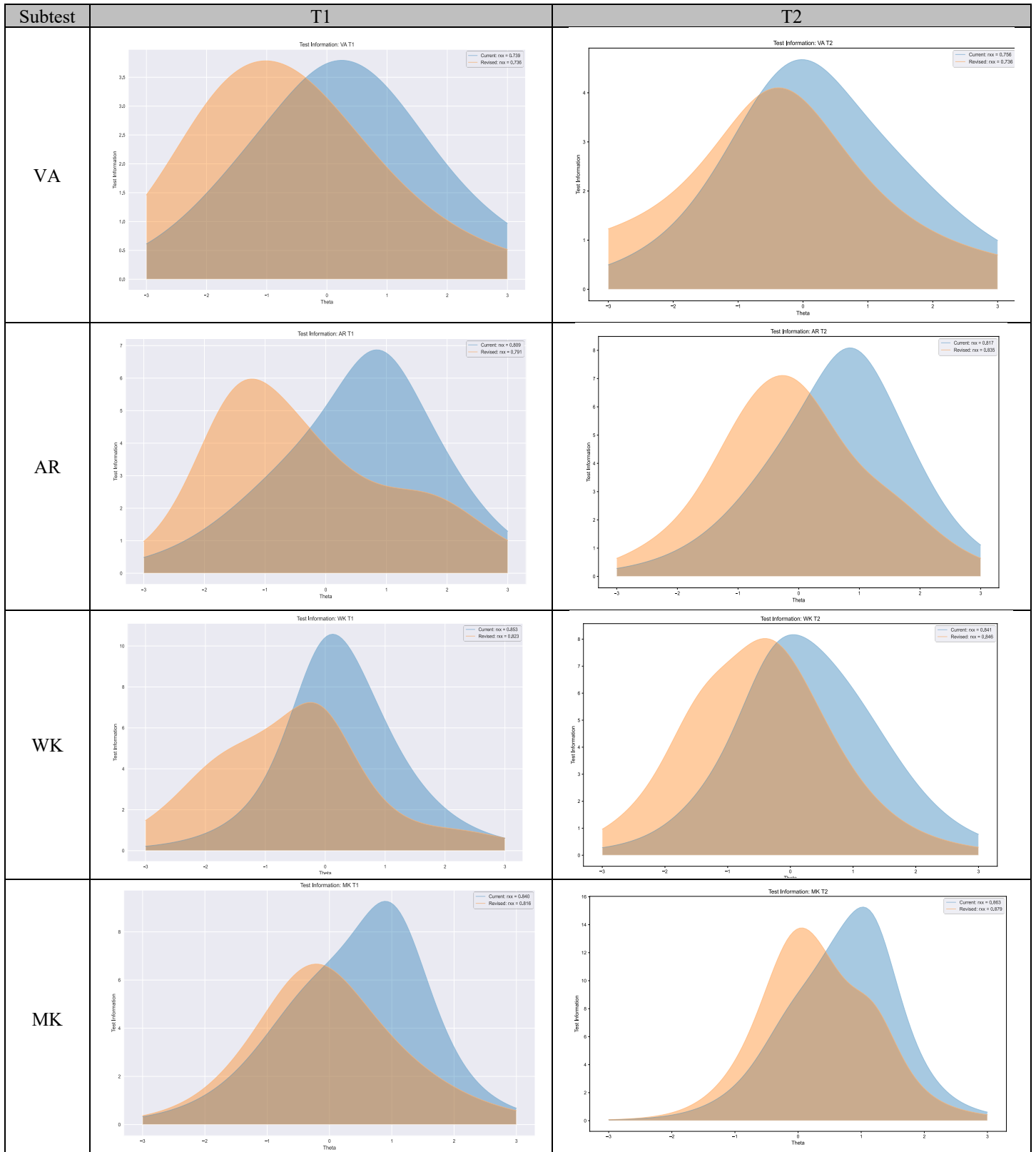
Table 8. Coefficient Alpha Reliability Estimates for Form Tv1 and Form Tv2 Subtests

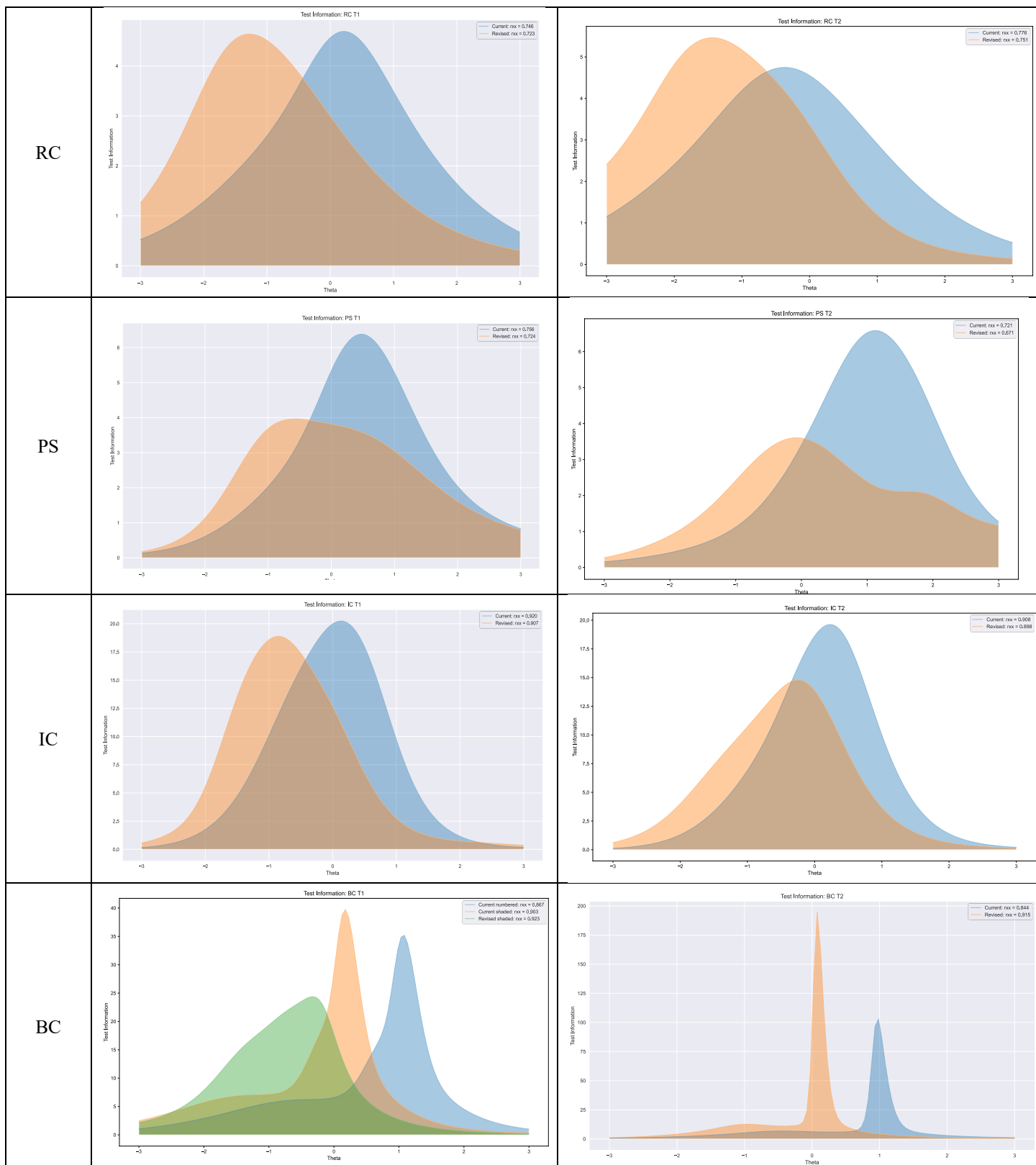
Subtest	Form T1 Reliability	Form Tv2 (A) Reliability	Difference	Form T2 Reliability	Form Tv2 (B) Reliability	Difference
			Tv2(A) – T1			Tv2(B) – T2
VA	.74	.74	.00	.76	.74	-.02
AR	.81	.79	-.02	.82	.84	.02
WK	.85	.82	-.03	.84	.85	.01
MK	.84	.82	-.02	.86	.88	.02
RC	.75	.72	-.02	.78	.75	-.03
PS	.76	.72	-.03	.72	.70	-.02
IC	.92	.91	-.01	.91	.90	-.01
BC	.87	.92	.06	.84	.92	.07
AI	.79	.75	-.03	.72	.71	-.02

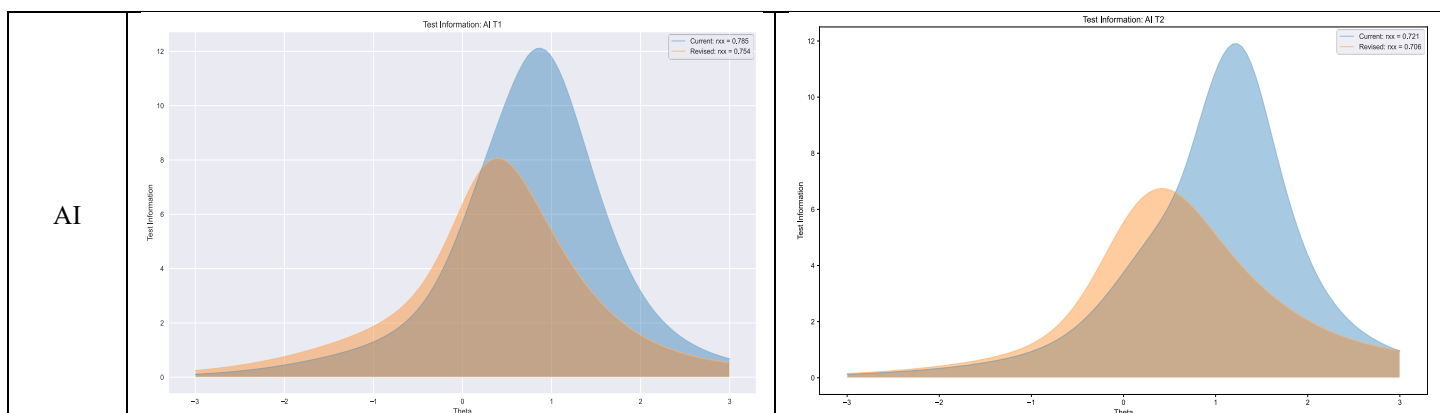
4.2.6 Test Information Function (TIF)

Indeed, as per our expectations, the TIFs for each subtest shifted closer to the latent ability distributions of non-male, non-White, non-Hispanic examinees. In addition, the amount of information was slightly reduced because of the intended shift in latent ability. Table 9 displays the differences in TIF for the current (Forms T1 and T2) and new (Forms Tv2 (A) and (B)) subtests.

Table 9. TIF







4.2.7 Subtest Composition

The form assembly process resulted in diversity improvements and maintained similar levels of reliability. As previously noted, a secondary goal of the development of AFOQT Form Tv2 was refreshing the content to increase the longevity of the assessments. To that end, an explicit goal was to ensure that the final forms included Form Tv1 and newly developed items. Table 10 shows the breakdown of the number of items comprising each subtest and how many items from Form Tv1 and newly developed were recommended.

Table 10. Final Item Counts by Source for Form Tv2

Subtest	Form Tv2 (A)			Form Tv2 (B)		
	Number of Recommended Items from Form T1	Number of Recommended New items	Total Items	Number of Recommended Items from Form T2	Number of Recommended New Items	Total Items
VA	15	10	25	10	15	25
AR	15	10	25	12	13	25
WK	10	15	25	10	15	25
MK	15	10	25	10	15	25
RC	0	25	25	0	25	25
PS	16	4	20	9	11	20
TR	25	15	40	25	15	40
IC	14	11	25	15	10	25
BC	17	13	30	17	13	30
AI	10	10	20	11	9	20

4.3 Objective 3: Equivalency/Equating Study

4.3.1 Equating Tables

With the exception of TR which used rate scoring, each AFOQT subtest's number right (NR) score was transformed to a standardized (t -score) with a M of 50 and SD of 10. To facilitate comparisons with previous AFOQT subtest scores, equating was accounted for in the t -score formula. Associated look-up tables for converting NR scores to t -scores are provided in Table 11 and Table 12

Table 11. Form Tv2(A) Number Right (NR) Converted to Equated *t*-Scores

VA		AR		WK		MK		RC		PS		IC		BC		AI	
NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score
0	3.72	0	9.83	0	4.14	0	11.39	0	-5.45	0	15.01	0	-5.3	0	15.1	0	20.47
1	6.34	1	12.17	1	6.58	1	13.57	1	-2.59	1	17.79	1	-0.86	1	16.73	1	23.11
2	8.97	2	14.51	2	9.02	2	15.74	2	0.26	2	20.56	2	3.59	2	18.35	2	25.75
3	11.6	3	16.85	3	11.46	3	17.91	3	3.12	3	23.33	3	8.04	3	19.97	3	28.39
4	14.22	4	19.19	4	13.9	4	20.08	4	5.98	4	26.1	4	12.49	4	21.6	4	31.03
5	16.85	5	21.53	5	16.34	5	22.25	5	8.84	5	28.88	5	16.93	5	23.22	5	33.67
6	19.47	6	23.87	6	18.79	6	24.42	6	11.7	6	31.65	6	21.38	6	24.85	6	36.31
7	22.1	7	26.2	7	21.23	7	26.6	7	14.56	7	34.42	7	25.83	7	26.47	7	38.95
8	24.73	8	28.54	8	23.67	8	28.77	8	17.42	8	37.19	8	30.27	8	28.1	8	41.59
9	27.35	9	30.88	9	26.11	9	30.94	9	20.28	9	39.96	9	34.72	9	29.72	9	44.23
10	29.98	10	33.22	10	28.55	10	33.11	10	23.14	10	42.74	10	39.17	10	31.34	10	46.87
11	32.61	11	35.56	11	30.99	11	35.28	11	26	11	45.51	11	43.61	11	32.97	11	49.51
12	35.23	12	37.9	12	33.44	12	37.45	12	28.86	12	48.28	12	48.06	12	34.59	12	52.15
13	37.86	13	40.24	13	35.88	13	39.63	13	31.72	13	51.05	13	52.51	13	36.22	13	54.79
14	40.48	14	42.58	14	38.32	14	41.8	14	34.57	14	53.82	14	56.95	14	37.84	14	57.43
15	43.11	15	44.92	15	40.76	15	43.97	15	37.43	15	56.6	15	61.4	15	39.47	15	60.07
16	45.74	16	47.26	16	43.2	16	46.14	16	40.29	16	59.37	16	65.85	16	41.09	16	62.71
17	48.36	17	49.6	17	45.64	17	48.31	17	43.15	17	62.14	17	70.3	17	42.72	17	65.35
18	50.99	18	51.94	18	48.09	18	50.49	18	46.01	18	64.91	18	74.74	18	44.34	18	67.99
19	53.62	19	54.28	19	50.53	19	52.66	19	48.87	19	67.68	19	79.19	19	45.96	19	70.63
20	56.24	20	56.62	20	52.97	20	54.83	20	51.73	20	70.46	20	83.64	20	47.59	20	73.27
21	58.87	21	58.96	21	55.41	21	57	21	54.59			21	88.08	21	49.21		
22	61.49	22	61.3	22	57.85	22	59.17	22	57.45			22	92.53	22	50.84		
23	64.12	23	63.63	23	60.29	23	61.34	23	60.31			23	96.98	23	52.46		
24	66.75	24	65.97	24	62.74	24	63.52	24	63.17			24	101.42	24	54.09		
25	69.37	25	68.31	25	65.18	25	65.69	25	66.03			25	105.87	25	55.71		
														26	57.33		
														27	58.96		
														28	60.58		
														29	62.21		
														30	63.83		

Note. NR = Number Right

Table 12. Form Tv2(B) Number Right (NR) Converted to Equated *t*-Scores

VA		AR		WK		MK		RC		PS		IC		BC		AI	
NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score	NR	<i>t</i> -score
0	8.09	0	16.59	0	14.45	0	15.33	0	-14.26	0	11.99	0	17.29	0	0.53	0	23.79
1	10.58	1	18.56	1	16.46	1	17.3	1	-11.19	1	14.73	1	19.07	1	2.63	1	26.43
2	13.08	2	20.54	2	18.47	2	19.27	2	-8.13	2	17.46	2	20.85	2	4.74	2	29.06
3	15.57	3	22.51	3	20.49	3	21.23	3	-5.07	3	20.2	3	22.63	3	6.85	3	31.7
4	18.07	4	24.48	4	22.5	4	23.2	4	-2.01	4	22.94	4	24.41	4	8.96	4	34.34
5	20.56	5	26.46	5	24.52	5	25.17	5	1.05	5	25.67	5	26.19	5	11.07	5	36.97
6	23.06	6	28.43	6	26.53	6	27.14	6	4.11	6	28.41	6	27.97	6	13.18	6	39.61
7	25.55	7	30.4	7	28.55	7	29.11	7	7.17	7	31.14	7	29.74	7	15.28	7	42.25
8	28.05	8	32.38	8	30.56	8	31.08	8	10.23	8	33.88	8	31.52	8	17.39	8	44.89
9	30.55	9	34.35	9	32.58	9	33.05	9	13.29	9	36.61	9	33.3	9	19.5	9	47.52
10	33.04	10	36.32	10	34.59	10	35.02	10	16.35	10	39.35	10	35.08	10	21.61	10	50.16
11	35.54	11	38.3	11	36.61	11	36.99	11	19.41	11	42.08	11	36.86	11	23.72	11	52.8
12	38.03	12	40.27	12	38.62	12	38.95	12	22.47	12	44.82	12	38.64	12	25.83	12	55.43
13	40.53	13	42.24	13	40.63	13	40.92	13	25.53	13	47.56	13	40.42	13	27.93	13	58.07
14	43.02	14	44.22	14	42.65	14	42.89	14	28.59	14	50.29	14	42.2	14	30.04	14	60.71
15	45.52	15	46.19	15	44.66	15	44.86	15	31.66	15	53.03	15	43.98	15	32.15	15	63.35
16	48.01	16	48.16	16	46.68	16	46.83	16	34.72	16	55.76	16	45.76	16	34.26	16	65.98
17	50.51	17	50.14	17	48.69	17	48.8	17	37.78	17	58.5	17	47.54	17	36.37	17	68.62
18	53	18	52.11	18	50.71	18	50.77	18	40.84	18	61.23	18	49.32	18	38.48	18	71.26
19	55.5	19	54.09	19	52.72	19	52.74	19	43.9	19	63.97	19	51.1	19	40.58	19	73.89
20	58	20	56.06	20	54.74	20	54.7	20	46.96	20	66.7	20	52.88	20	42.69	20	76.53
21	60.49	21	58.03	21	56.75	21	56.67	21	50.02			21	54.66	21	44.8		
22	62.99	22	60.01	22	58.77	22	58.64	22	53.08			22	56.44	22	46.91		
23	65.48	23	61.98	23	60.78	23	60.61	23	56.14			23	58.22	23	49.02		
24	67.98	24	63.95	24	62.79	24	62.58	24	59.2			24	60	24	51.13		
25	70.47	25	65.93	25	64.81	25	64.55	25	62.26			25	61.78	25	53.23		
														26	55.34		
														27	57.45		
														28	59.56		
														29	61.67		
														30	63.78		

Note. NR = Number Right

4.3.2 Equating Approach for TR

As noted in Section 2, a different approach was required to equate the TR subtest since the goals were to (1) preserve the meaning of scores to previous versions of AFOQT and (2) equate the web-based version to the previous, paper-based version. The use of rate scoring was a way to meet these goals.

The data to determine rate scoring was derived from data collected with officer candidates at OTS and SOS. For both OTS and SOS, each participant completed one of two subtests: Form T1 or T2 in paper-based format or Form Tv2 in web-based format. For Tv2 (A), the cleaned sample size ($N = 637$) provided the basis for analysis. For Tv2 (B), the cleaned sample was $N = 610$. For Tv2 (A), it should be noted that due to a system glitch, the time limit for Form Tv2 (A) was set to 7 minutes in the OTS administration but was untimed in the SOS administration. See Table 13.

Table 13. Sample Sizes associated with TR Data Collection

Source	Tv2 (A)			Tv2 (B)		
	Paper-Based	Web-Based	Total	Paper-Based	Web-Based	Total
Total	336	301	637	291	319	610

To determine the rate score that equated Form Tv2 web-based to Form Tv1 paper-based, we started by calculating rate scores for the participants who completed the web-based version of the assessment (separately for Tv2 [A] and Tv2 [B]), which was the corrected proportion right divided by the mean response time. To determine whether the SOS and OTS groups could be combined, we evaluated whether significant differences existed in means and variances. No statistically significant differences in means or variances were found. Then, we calculated the NC scores for the participants who completed the paper-based assessment. Again, the differences in means and variances were evaluated, and no statistically significant differences were found. Given the similarity in means and variances, data from the two groups were combined to create a master data set comprising data from two conditions (paper-based and web-based administrations). Linear equating was then conducted by converting Form T1v1 paper-based number right scores to t -scores using the operational T1 scoring transformation. The linear transformation of Tv2 (A) rate scores was calculated that produced the same t -score M and SD as the paper-based group.

The resulting rate score distribution was compared across the SOS and OTS groups. The difference between the groups was compared and found not to be statistically significantly different. See Table 14 (web-based administration) and Table 15 (paper-based administration) for details. The same procedure was used for equating Tv2(B) to T2.

Table 14. Rate Score Comparisons between OTS and SOS Subsamples – Web-Based Administration

	OTS	SOS	t -test	p -value
M	4.45	4.44	.26	.80
SD	1.29	1.31	.97	.57

Table 15. Number Right Score Comparisons between OTS and SOS Subsamples – Paper-Based Administration

	OTS	SOS	<i>t</i> -test	<i>p</i> -value
<i>M</i>	25.52	25.58	-.09	.93
<i>SD</i>	6.35	6.66	.91	.74

Based on these analyses and the comparability of the data across sub-samples, the data were combined across the OTS and SOS groups for both the paper-based administrations (based on NR scoring) and web-based administrations (based on rate scores) to increase the accuracy of the equating transformation. In addition, analysis of the sensitivity of the rate score distributions for the timed (SOS) and untimed (OTS) administrations showed that the groups scored comparably. The equating transformation was determined based on the combined SOS and OTS samples.

The equating formulas are:

- Tv2(A): $11.68671 + 8.06433 * \text{rate_score_TR}$
- Tv2(B): $13.26235 + 8.72038 * \text{rate_score_TR}$

4.4 Objective 4: Develop Two Forms of AFOQT Form Tv2 and Support New Composite Scoring

4.4.1 Composite Equating Results

It was determined to implement updated scoring for the Pilot, CSO, and ABM composites that included selected facets from the SDI-O. USAF desired to preserve the ability to equate the new composite scores to the legacy composite scores. Given the dissimilarity between legacy composites (which did not include SDI-O facets) and new composites (which included SDI-O facets), it was not assumed or expected that the composites would be equatable. Rather, evaluation of concordance was proposed for feasibility analysis.

Composite Content Review

The first step involved a review of content and test specification similarity. Per the tests and weights specified in Woolley et al., (2023), it was noted that the new composites add additional constructs, which limit the ability to equate and exchange scores across test forms.

Inter-Correlations, Reliability, and Validity

Next, the inter-correlations, reliability, and validity of the Form Tv1 and Form Tv2 composites were examined. The inter-correlation between the two sets of composites was found to be moderately high, with the correlation coefficient ranging from .80 (for ABM) to .88 (for Pilot). The reduction in uncertainty (RiU; Dorans, 2004) an index of alienation, is a measure of statistical uncertainty about a score that remains after inclusion of information from another score. Per Dorans (2004), “If a predictor cannot reduce uncertainty by at least 50%, it is unlikely that it can serve as a valid surrogate, via concordance or equating, for the score being predicted” (p. 240). As

RiU is an extension of the correlation, Dorans also indicated that correlations in the low .80s are too low to produce linkages with uncertainty reduction of at least 50%.

The results are shown in Table 16. For the Pilot composite, the correlation was .88 and the RiU was .53, indicating that composite scores on Form Tv2 are 53% likely to be predicted from Form Tv1 composite scores, and as such, achieved the minimum of 50% uncertainty reduction. In addition, the correlation was minimally adequate for using the Form Tv1 and Form Tv2 composites as surrogates. For the CSO composite, however, the correlation was .82 and it was 43% likely that the Form Tv2 composite scores can be predicted by Form Tv1 composite scores, which did not meet the 50% threshold provided by Dorans (2004). The correlation was lower than required to use the composites as surrogates. Finally, for the ABM composite, the correlation was .80 and the RiU was .40, suggesting that it was 40% likely that the Form Tv2 composite scores can be predicted by Form Tv1 composite scores, which also did not meet the 50% threshold provided by Dorans. The correlation was also lower than required to use the composites as surrogates. This evidence did not support a basis for score concordance.

Table 16. Correlation and Reduction in Uncertainty Values between Form Tv1 and Form Tv2 Composites

Career Field	r_{xy}	RiU
Pilot	.88	.53
CSO	.82	.43
ABM	.80	.40

The reliability estimates of the Form Tv1 and Form Tv2 composites were calculated and are summarized in Table 17. Coefficient alpha (Cronbach, 1951), a measure of the internal consistency, is one of the most frequently used coefficients for measuring the reliability of a test. Lord and Novick (1968) showed that all items must be tau-equivalent for coefficient alpha to be an unbiased estimator of the reliability. Items are tau-equivalent when the difference between the true scores for any pair of items is a constant and the items have equal true score variance. However, a composite score made up of multiple constructs generally does not meet these criteria. Instead, a composite can be assumed to be composed of stratified layers of items, with items in each layer being assumed to be unidimensional. Thus, the reliability of the composite score can be calculated based on the reliabilities of the strata, each of which is treated as a single subtest (He, 2009). Additionally, among other measures of composite reliability, Kamata and colleagues (2003) showed that stratified alpha generally performs the best of the various composite measures. Therefore, stratified alpha was used to estimate the reliability of the Form Tv1 and Form Tv2 composites.

Table 17. Composite Reliability Estimates for Form Tv1 and Form Tv2

Career Field	Form Tv1 Composite Reliability	Form Tv2 Composite Reliability
Pilot	.94	.90
CSO	.92	.91
ABM	.95	.91

Reliability for Form Tv1 composites were found to be slightly higher than Form Tv2, which was not surprising given that Form Tv1 composites involve exclusively cognitive subtests (which tend to demonstrate positive manifold), whereas the Form Tv2 composites include a mixture of cognitive and non-cognitive subtests. The lower reliability for Form Tv2 was expected to potentially influence validity estimates.

The criterion-related validity of the Form Tv1 and Form Tv2 composites was then compared. Validity estimates were determined in a separate effort (see Woolley et al., 2023) and utilize a variety of criteria, including training outcome measures collected at Initial Flight Training (IFT) and Supervised Undergraduate Pilot Training (SUPT) Primary and Advanced. The validity estimates are shown in Table 18.

The validity of the Form Tv2 composites was generally found to be higher than those for Form Tv1, but it was noted to be lower for the Pilot composite in the prediction of IFT Merit Assignment Selection System (MASS). This may have been somewhat due to the lower composite reliability estimate for Form Tv2 pilot composite compared to Form Tv1. Nonetheless, the lower validity for the IFT MASS criteria challenged the finding of concordance. Validity estimates, expressed as the percentage of variance accounted for in the criterion measure by the composite predictor, are shown in Table 18.

Table 18. Criterion-Related Validity Estimates for Form Tv1 and Form Tv2 Composites

Career Field	Form Tv1 Composite Validity Estimates	Form Tv2 Composite Validity Estimates
Pilot	IFT Mass $r^2=13.7$ SUPT Primary $r^2=2.2$ SUPT Advanced $r^2=3.69$	IFT Mass $r^2=4.92$ SUPT Primary $r^2=8.69$ SUPT Advanced $r^2=3.69$
CSO	IFT Mass $r^2=2.79$ CSO Primary $r^2=4.33$ SUPT Advanced $r^2=5.51$	IFT Mass $r^2=6.41$ CSO Primary $r^2=8.09$ SUPT Advanced $r^2=6.82$
ABM	ABM Primary $r^2=8.31$	ABM Primary $r^2=15.3$

Group Invariance

Lack of score invariance across groups complicates the use of comparison tables to compare scores derived using different composites. For instance, if corresponding scores between Form Tv1 and Form Tv2 differ for males and females, this would introduce challenges in score use and interpretation and may introduce legal challenges. To evaluate whether scores are invariant across groups, SGDs and passing rates were calculated using Form Tv1 data consisting of more than 70,000 examinees. For comparison purposes, the passing rate was based on the 25th percentile for qualification. Results for the Pilot composite are shown in Table 19. Subgroup differences were calculated as the minority group score – the majority group score. Negative values indicate lower scores for the minority group. A positive Difference score indicates that SGDs were smaller for Form Tv2 than for Form Tv1.

Table 19. Subgroup Differences and Passing Rate Comparison – Pilot Composite

Subgroups	Form Tv1 Subgroup Difference	Form Tv2 Subgroup Difference	Difference	Form Tv1 Composite Passing Rate	Form Tv2 Composite Passing Rate	Difference
Female	-.88	-.84	+.04	48.75	49.84	+1.08
Black/African American	-1.38	-1.09	+.29	34.54	37.51	+2.97
Hispanic/Latino	-.43	-.57	-.14	59.34	60.27	+.93
Asian	-.32	-.30	+.02	71.97	70.95	-1.02
Native Hawaiian/Pacific Islander	-.83	-.50	+.33	65.06	64.11	-.95
American Indian/Alaskan Native	-.74	-.45	+.29	65.85	64.29	-1.56

In comparing Form Tv1 and Form Tv2 SGD, modest reductions in SGD were noted for some minority groups (Black, Native Hawaiian/Pacific Islander, American Indian/Alaskan Native) but a slight increase for Hispanic/Latino. Similarly, modest increases in passing rates were noted for some groups (Female, Black) but modest decreases for others (Asian, Native Hawaiian/Pacific Islander, American Indian/Alaskan Native). The combination of increases and decreases means that the score linking procedure would advantage some groups but disadvantage others.

Results for the CSO composite are presented in Table 20. In comparing Form Tv1 and Form Tv2 SGD, near zero reduction or increase in SGD was noted for most minority groups but a slight increase in SGD for Hispanic/Latino. A modest increase in passing rates was noted for some groups (Female, Black/African American, Native Hawaiian/Pacific Islander) but modest decreases for others (American Indian/Alaskan Native). The combination of increases and decreases means that score linking procedures would advantage some groups but disadvantage others.

Table 20. Subgroup Differences and Passing Rate Comparison – CSO Composite

Subgroups	Form Tv1 Subgroup Difference	Form Tv2 Subgroup Difference	Difference	Form Tv1 Composite Passing Rate	Form Tv2 Composite Passing Rate	Difference
Female	-.42	-.34	+.08	66.34	68.04	+1.71
Black/African American	-1.30	-1.34	-.04	44.03	44.96	+.93
Hispanic/Latino	-.41	-.86	-.45	65.70	65.69	-.01
Asian	-.28	-.26	+.02	77.64	77.85	+.21
Native Hawaiian/Pacific Islander	-.76	-.69	+.07	72.01	74.42	+2.41
American Indian/Alaskan Native	-.72	-.55	+.17	72.31	71.57	-.75

Table 21 summarizes the results for the ABM composite. In comparing Form Tv1 and Form Tv2, modest reductions in SGDs were noted for several minority groups (Female, Black/African American) but increases for others (Hispanic/Latino). Large or modest increases in passing rates were also noted for some groups (Female, Black) but decreases for others (Asian). As such, the combination of increases/decreases means that score linking procedures would advantage some groups but disadvantage others.

Table 21. Subgroup Differences and Passing Rate Comparison – ABM Composite

Subgroups	Form Tv1 Subgroup Difference	Form Tv2 Subgroup Difference	Difference	Form Tv1 Composite Passing Rate	Form Tv2 Composite Passing Rate	Difference
Female	-.63	-.33	+.30	56.29	63.90	+7.61
Black/African American	-1.37	-1.02	+.35	36.40	47.70	+11.29
Hispanic/Latino	-.44	-.74	-.30	60.36	61.84	+1.48
Asian	-.19	-.30	-.11	74.93	71.14	-3.79
Native Hawaiian/Pacific Islander	-.81	-.63	+.18	66.69	66.42	-.27
American Indian/Alaskan Native	-.74	-.54	+.20	66.30	66.23	-.07

Effect of Faking on Composite Scores

The inclusion of the SDI-O facets in Form Tv2 composite scores was evaluated to determine the role that potential for faking/impression management may have on composite scores. The SDI-O uses a Likert-type response scale which was thought to be vulnerable to impression management as a function of response format. In addition, the communication to examinees (e.g., AFOQT Form Tv1 pamphlet) has instructed examinees that the SDI-O has been unscored in classification composites. If updated to disclose the scored nature of the SDI-O in operational composites, it was thought that SDI-O scores may increase as candidate motivation to score highly increases. However, it was recognized that the use of non-linear scoring in the new composites (see Woolley et al., 2023) may mitigate the potential effects of faking for some facets and composites.

To evaluate whether faking may complicate potential score concordance for Form Tv1 and Form Tv2 composites, we started by examining the specific facet scores intended for Form Tv2 composites and the associated item content. Multiple reviewers estimated the degree of “fakability” associated with the items in a given facet. The expectations for fakability were used to inform a method for simulating the effect of faking on composite scores. The expectations for fakability, the resulting method for simulating the effect of faking, and the result are shown in Table 22.

The results of the faking simulation showed the effect of faking in the manner outlined may have substantial influence on composite scores for Pilot, CSO, and ABM. Furthermore, the weights and non-linear effects in the composite scoring formula may not generalize if the response patterns

change in operational administration due to faking. Once it is disclosed to examinees that the SDI-O is a scored component of the AFOQT, we recommend that operational test data be reviewed to determine if the composite scoring models generalize under this condition.

Table 22. Faking Evaluation of SDI-O Facets and Effects on Composite Score Distributions

Career Field	Predicted Level of Fakability based on Item Content Review	Method for Simulating the effect of Faking on Composite Scores
Pilot	Temperamental: very fakable Worry: somewhat fakable Independent: less fakable Cultured: less fakable	Simulating faking for Temperamental only (raw score =1), 40% of candidates are expected to gain 10 or more percentile points on the Pilot composite
CSO	Creativity: very fakable Reflective: slightly fakable Envious: very fakable	Simulating faking for Creativity (raw score = 5) and Envious (raw score = 1), 22% of candidates are expected to gain 10 or more percentile points on the ABM composite
ABM	Reflective: slightly fakable Self-Discipline: very fakable Scientific Interest: less fakable Envious: very fakable	Simulating faking for Envious (raw score = 1) only, 5% of candidates are expected to gain 10 or more percentile points on the ABM composite Simulating faking for Self-Discipline (raw score = 5) and Envious (raw score = 1), 22% of candidates are expected to gain 10 or more percentile points on the ABM composite

Conclusions

In conclusion, Form Tv1 and Form Tv2 composites measure different constructs, which does not permit score equating. The correlation and reduction in uncertainty between the existing and new Pilot composites is minimally adequate for using the new and existing composites as surrogates, but a decrease in validity is expected. The correlations and reduction in uncertainty between the existing and new ABM and CSO composites are not minimally adequate for using the Form Tv1 and Form Tv2 composites as surrogates. Scores are not invariant across subpopulations so a single linking function of Form Tv2 to Form Tv1 scores would advantage some groups and disadvantage other groups. The effect of faking may substantially change Form Tv2 scores from original expectations. Overall, requirements for concordance are not met.

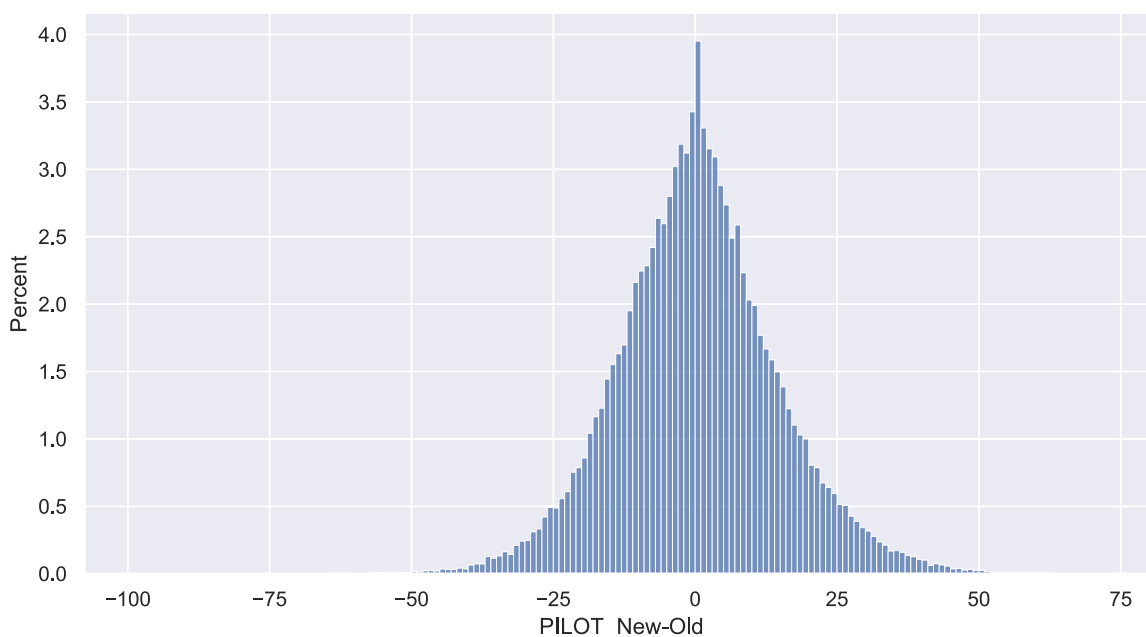
Implementation Options

We developed two options for composite scoring in light of the concordance evaluation. The first option was to not link new and existing composites as scores are not exchangeable and evidence of score concordance is not met. In this scenario, implementing new composites was seen as dependent on new norms to support the new composites, waiting until Form Tv1 is retired, and would require all candidates to retest with the knowledge that SDI-O facet scores are used operationally. The second option was to implement the Form Tv1 (cognitive only) and Form Tv2 (cognitive + SDI-O) composites based on equipercentile equating such that candidates can pass on either score (through the use of existing cut-scores). The operational score would be the maximum of both and would not require the scores to be exchanged. It was noted that this option would likely result in a slightly higher passing rate. The benefits of this option are that it would allow for further review of the equating formula and review of scored use of the SDI-O in composites as response patterns may change once it's known that the SDI-O is a scored test, it would allow Form Tv2 to

be implemented for new paper-based testing to allow for equal access to new content and scoring, and would allow previous candidate scores remain as-is.

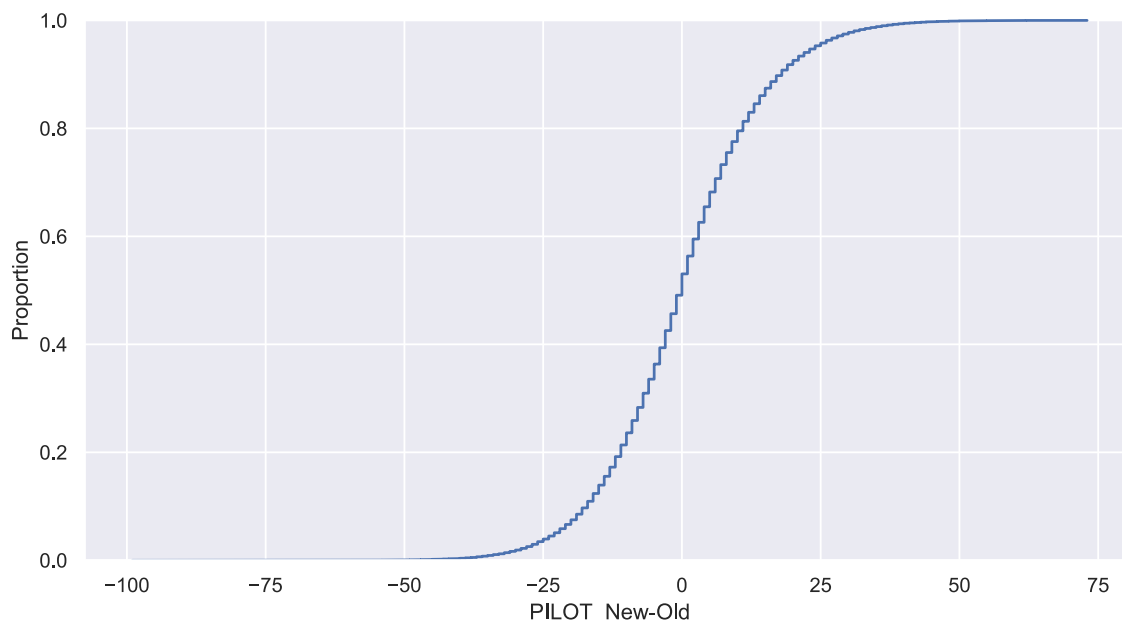
As a follow-up to the concordance feasibility analysis, the Form Tv1 and Form Tv2 Pilot, CSO, and ABM composite scores were compared. The *t*-scores were transformed to percentile scores. For Form Tv2, weighted scores were transformed to the percentile score scale using an equipercentile approach. As a result, any given score (from 1 to 99) provides the same qualification rate across the two sets of composites. The score comparison data were derived from the Form Tv1 officer candidate data. The objective of this analysis was to understand the degree and frequency of score differences between the Form Tv1 and Form Tv2 composites.

For the Pilot composite, score differences were found to be normally distributed with most score differences falling between -10 and +10 percentile points (see Figure 1), though it was noted that differences of 15 or more points were observed in 20% of the cases. As can be seen in Figure 2, the cumulative density function shows that almost all scores fall within a -25 to +25 point difference. It was noted that scores are as likely to increase as decrease.



Note: Pilot_New-Old = Form Tv1 composite minus Form Tv2 composite

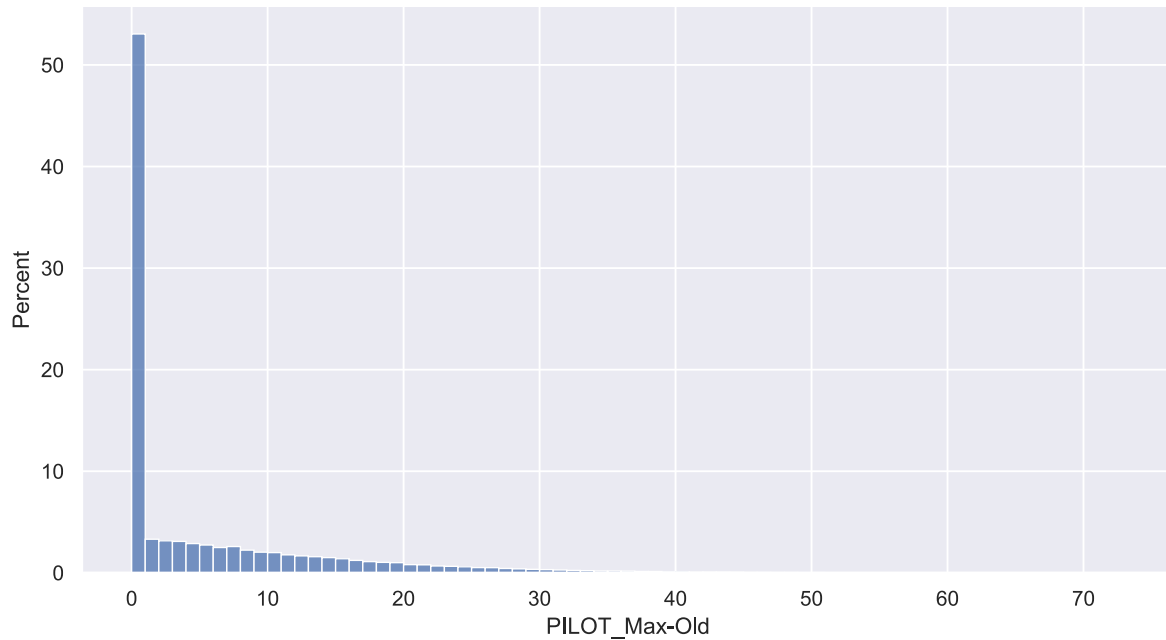
Figure 1. Distribution of Score Differences for Form Tv1 and Form Tv2 Composites – Pilot



Note: Pilot_New-Old = Form Tv1 composite minus Form Tv2 composite

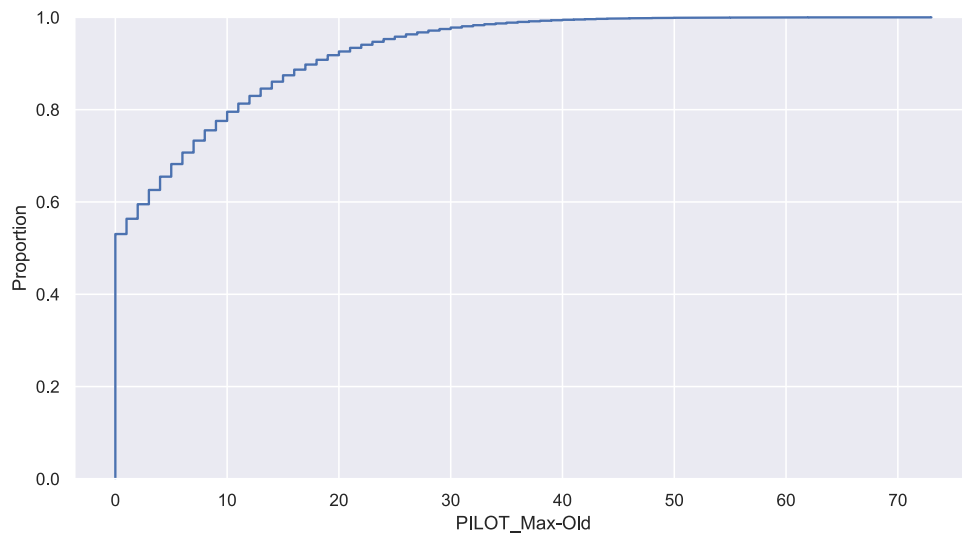
Figure 2. Cumulative Score Difference Density for Form Tv1 and Form Tv2 Composites – Pilot

As noted above, the dual-composite approach calls for the maximum score as the final score of record. For the majority of candidates, the Form Tv1 Pilot composite score was found to be the maximum score. For the minority of candidates for whom the Form Tv2 score was the maximum, most score differences are within 10 percentile points. Results are shown in Figure 3 and in Figure 4. Figure 3 shows the magnitude of score differences between the maximum Pilot composite score and the Form Tv1 composite score. Figure 4 shows the cumulative distribution of score differences between the two composite scores.



Note: Pilot_Max-Old = Maximum composite minus Form Tv1 composite

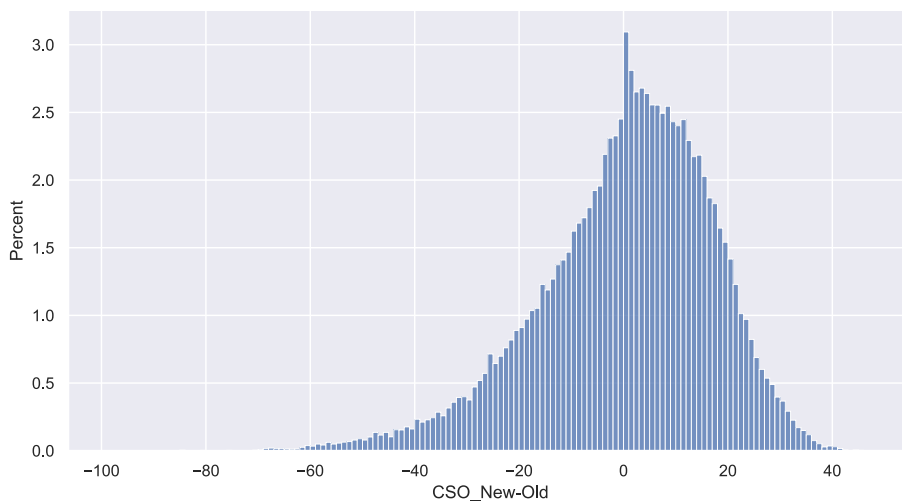
Figure 3. Distribution of the Score Difference Between Maximum Composite and Form Tv1 Composite – Pilot



Note: Pilot_Max-Old = Maximum composite minus Form Tv1 composite

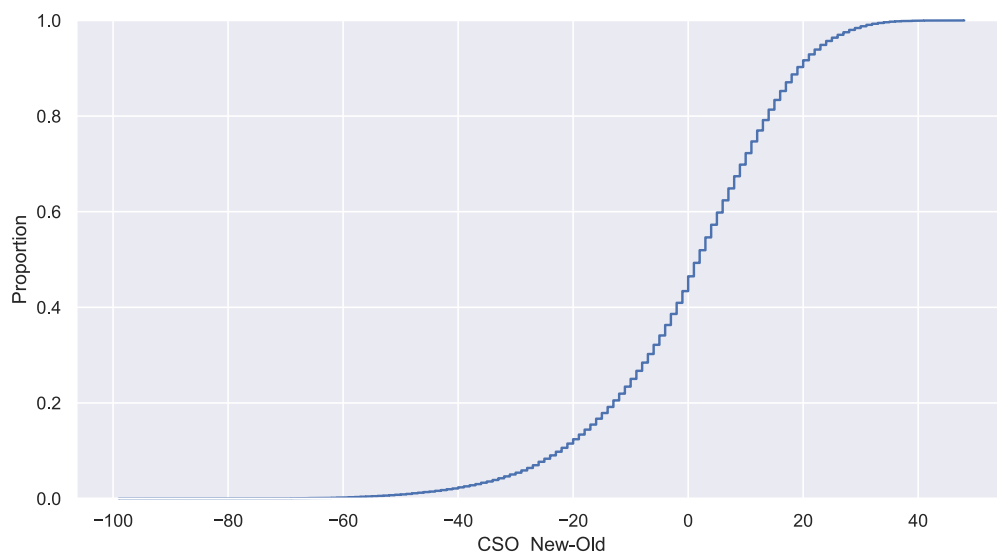
Figure 4. Cumulative Score Difference Density for Maximum Composite Score and Form Tv1Composite Score – Pilot

Similar results were observed for CSO. As shown in Figure 5, score differences were found to be somewhat negatively skewed with most differences between -15 and +15 percentile points. Scores were slightly less likely to decrease than increase. Figure 6 cumulative density of score differences and that almost all score differences fall between -40 and +20.



Note: CSO_New-Old = Form Tv1 composite minus Form Tv2 composite

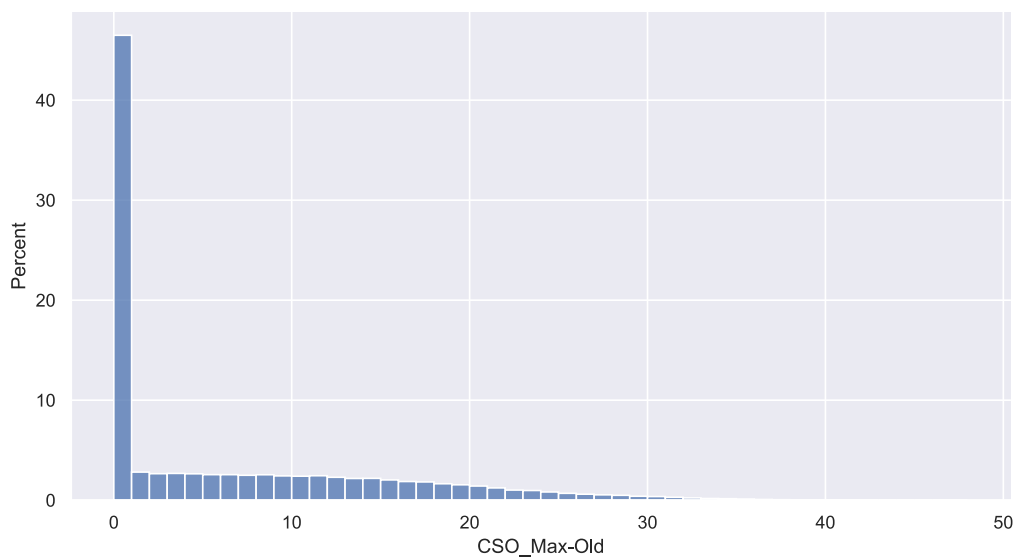
Figure 5. Distribution of Score Differences for Form Tv1 and Form Tv2 Composites – CSO



Note: CSO_New-Old = Form Tv1 composite minus Form Tv2 composite

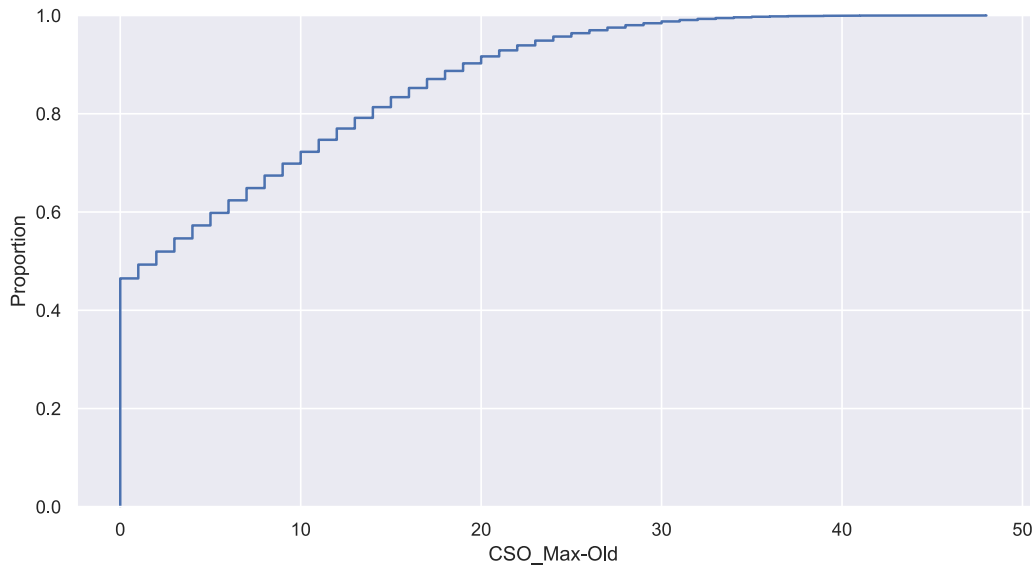
Figure 6. Cumulative Score Difference Density for Form Tv1 and Form Tv2 Composites – CSO

When comparing the maximum score to Form Tv1 scores, for the largest portion of candidates (though not a majority), the Form Tv1 composite score was the maximum score. For candidates for whom the Form Tv2 score is the maximum score, most score differences are within 20 percentile points. See Figure 7 and Figure 8 for details. Figure 7 shows the magnitude of score differences between the maximum CSO composite score and the Form Tv1 composite score. Figure 8 shows the cumulative distribution of score differences between the two composite scores.



Note: CSO_Max-Old = Maximum composite minus Form Tv1 composite

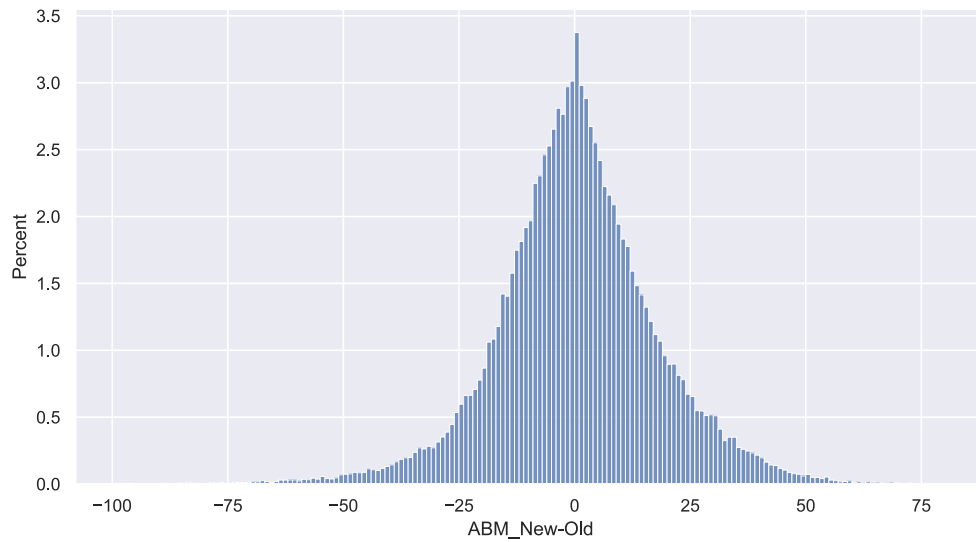
Figure 7. Distribution of Score Difference for Maximum and Form Tv1 Composite – CSO



Note: CSO_Max-Old = Maximum composite minus Form Tv1 composite

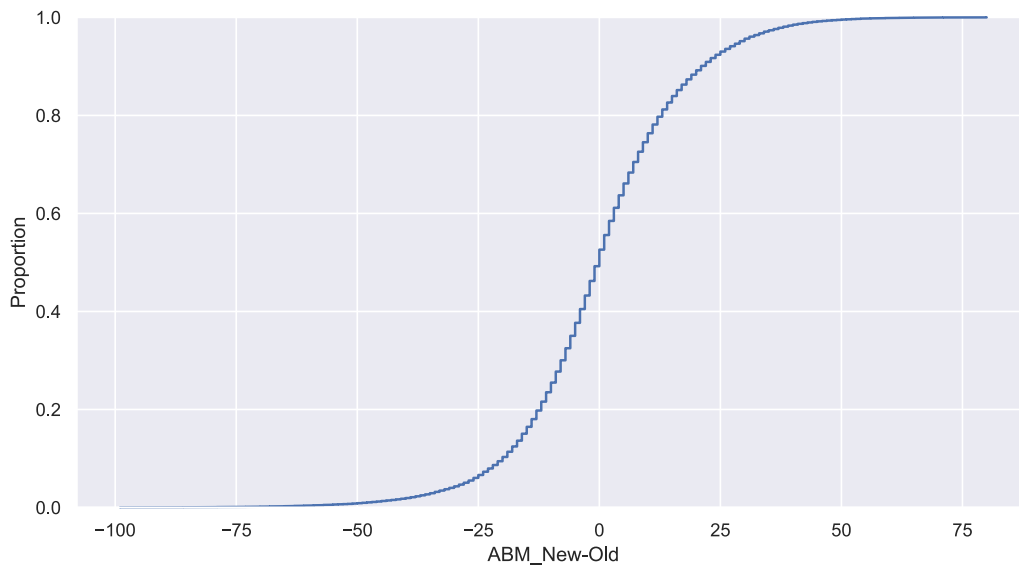
Figure 8. Cumulative Score Difference Density for Maximum and Form Tv1 Composite Score – CSO

Similar results were observed for ABM. As shown in Figure 9, score differences were found to be normally distributed, with most differences between -10 and +10 percentile points (though differences exceeding 15 points were found in 20% of the cases). Scores were found to be equally likely to increase as decrease. Figure 10 shows the cumulative density of score differences and that most score differences fall between -25 and +25.



Note: ABM_New-Old = Form Tv1 composite minus Form Tv2 composite

Figure 9. Distribution of Score Differences for Form Tv1 and Form Tv2 Composites – ABM

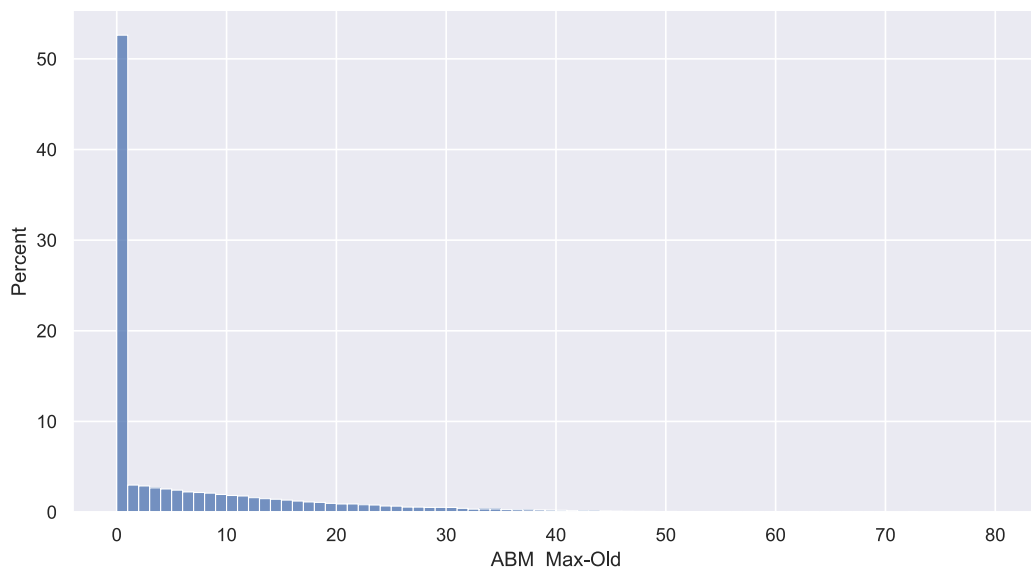


Note: ABM_New-Old = Form Tv1 composite minus Form Tv2 composite

Figure 10. Cumulative Score Difference Density for Form Tv1 and Form Tv2 Composites – ABM

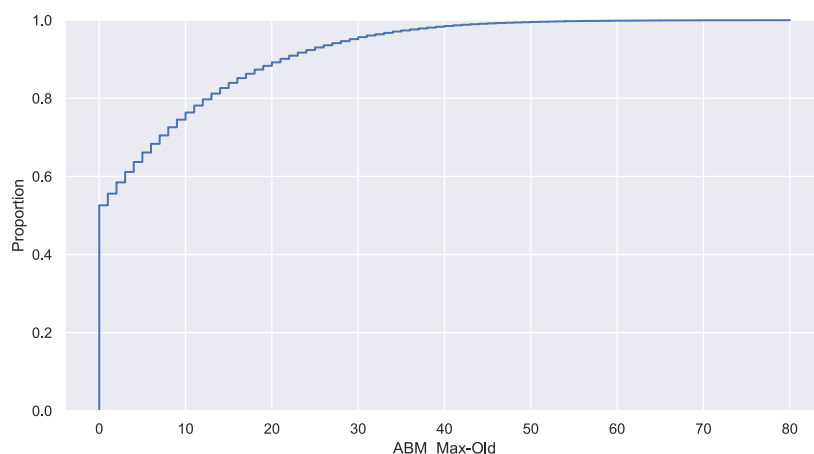
For the majority of candidates, the Form Tv1 score was found to be the maximum score. For the minority of candidates for whom the Form Tv2 score was the maximum, most score differences

were found to be within 10 percentile points, though it was noted that differences of 20 percentile points were found for a non-trivial number of candidates. See Figure 11 and Figure 12 for details. Figure 11 shows the magnitude of score differences between the maximum ABM composite score and the Form Tv1 composite score. Figure 12 shows the cumulative distribution of score differences between the two composite scores.



Note: ABM_Max-Old = Maximum composite minus Form Tv1 composite

Figure11. Distribution of Score Difference for Maximum and Form Tv1 Composite Score – ABM



Note: ABM_Max-Old = Maximum composite minus Form Tv1 composite

Figure 11. Cumulative Score Difference Density for Maximum and Form Tv1 Composite Score – ABM

Overall, it was found that if Form Tv2 composite scores are used exclusively, it would lead to substantially different rank ordering of candidates, as many candidates will experience significant score increases and decreases compared to Form Tv1 composites. Differences exceeding 10 percentile points were not found to be uncommon. Furthermore, there was no uniform increase noted for Form Tv2 composites due to the use of the equipercentile transformation applied to the Form Tv2 composites to place scores on the Form Tv1 composite score scale. The maximum scoring approach was found to lead to fewer rank order changes compared to Form Tv2 composites on their own. In general, approximately 50% of candidates were found to have the same score and the maximum scoring approach will not disadvantage any candidates. Based on a review of all findings, the Air Force opted to implement the second option. Composite scoring procedures were updated to account for the dual-composite scoring approach.

4.5 Discussion

The results showed that scale development methodology (in which test forms were generated and evaluated to determine which subset of Form Tv1 and Form Tv2 items resulted in the greatest potential for diversity which minimizing the decrement in reliability) produced test forms that were able to achieve the goals for Form Tv2. Across the subtests, the Form Tv2 versions showed that the gap between qualified candidates and applicants, by subgroup, was reduced compared to that gap that existed for Form Tv1 subtests. The subtests demonstrated test information functions that were closer to the latent ability distributions of minority subgroups. The decrement in reliability estimates was minimal across subtests, preserving a high degree of psychometric precision in test scores.

Through a series of equating procedures, Form Tv2 subtests were equated to Form Tv1 scores, preserving the meaning of scores across versions. Composite equating was evaluated to determine if scores on Form Tv2 composites (Pilot, CSO, and ABM, which introduce the SDI-O as a score component of AFOQT) could be equated to Form T. The dissimilarity in the content and constructs was noted as a limitation to equating, and various data points were evaluated to determine if scores could be linked. While some evidence supported score concordance, some did not. As such, composite scores could not be considered exchangeable. Instead, we recommended that a dual composite approach be implemented, in which Form Tv1 and Form Tv2 composite scores are simultaneously computed, and the maximum score retained. This approach enabled ongoing review of Form Tv2 composites to determine if they function as intended upon implementation with operational examinees.

5.0 CONCLUSION

The goals for the development of AFOQT Form Tv2 were to (1) refresh a portion of the subtest assessment content to extend the shelf life of the assessment until a complete refresh of the AFOQT can be conducted and (2) determine which subsets of assessment items yielded maintenance of psychometric score precision while also maximizing the potential for lower subgroup differences and greater candidate diversity. These goals were investigated and executed across four objectives, including: (1) replacement of problematic AFOQT Form Tv1 items, (2) data collection, item analysis, and scale development, (3) equating analysis to Form T, and (4) development of two forms and evaluation of new composites for Pilot, CSO, and ABM classification. Across the

objectives, the results indicated that Form Tv2 subtests achieved greater potential for diversity compared to Form Tv1 subtests while maintaining similar levels of psychometric precision.

The use of Form Tv2 has implications for operational administration of the AFOQT. The new subtests, once implemented on web-based testing platforms, will enable the selection and classification of prospective Air Force officers that more closely resemble the demographic characteristics of a more diverse set of officer candidates compared to previous versions of the AFOQT. The implementation of Form Tv2 will enable the Air Force to respond to the D&I strategic plan published in 2012. The results point to preservation of psychometric quality while reducing systematic differences in scores across minority demographic groups such as race, ethnicity, and gender. Going forward, we recommend the implementation of Form Tv2 subtests and super-scoring for two sets of composites for Pilot, CSO, and ABM. This approach will enable continued review of Form Tv2 composites to determine whether disclosure of SDI-O scoring in composites changes examinee response patterns.

6.0 REFERENCES

- Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on Black–White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology*, 66, 91–126.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Dorans, N. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28, 227-246.
- He, Q. (2009). *Estimating the reliability of composite scores*. Technical report. The Office of Qualifications and Examinations Regulation.
- Kamata, A., Turhan, A., & Darandari, E. (2003). Estimating reliability for multidimensional composite scale scores. In *annual meeting of American Educational Research Association*, Chicago, IL.
- 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*. AFRL-RH-WP-TR-2022-0038. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Systems Directorate Airman Biosciences Division, Performance Optimization Branch.
- Lord, F. M. & Novick, M. R. (1968) *Statistical Theories of Mental Test Scores*. Addison-Wesley, Menlo Park.
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54, 1643-1662.
- Segall, D. O., Moreno, K. E., Bloxom, B., Hetter, R. D. (1997). Psychometric procedures for administering CAT-ASVAB. In W. A. Sands, B. K. Waters, and J. R. McBride (eds.), *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.
- Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40(4), 331-352.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of DIF using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum, 67-113.
- 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., & 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: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
- Woolley, M. R., Walsh, J. L., Carretta, T. R., Mouton, A., & Deregla, A. (2023). *Development and psychometric evaluation of predictive success models for US Air Force rated career fields*, AFRL-RH-WP-TR-2023-0007. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Skokie, IL: Scientific Software International, Inc.

APPENDIX. CLASSICAL TEST THEORY ITEM-LEVEL ANALYSES

This appendix describes the methods and results of the CTT item-level analyses performed on the new items that were developed as part of the AFOQT Form Tv2 effort. The psychometrics for the new items as well as for the anchor items that were included as part of data collections are available in the Attachment 1 “CTT Parameters.xlsx.” Note that as part of the data collections, attention check and calculator check items were also included, but they were used in data cleaning phase and therefore were not analyzed here.

Methods

CTT is a commonly used and widely accepted item-level and scale-level methodology (Furr & Bacharach, 2008; Nunnally & Bernstein, 1994). CTT rests on several assumptions, including (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 the CTT 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 for the fact that its statistics and parameters are sample- and test-dependent which makes the comparability among the test administrations difficult. Second, participants' 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 by IRT. For the purposes of the current project, IRT was used to evaluate the new items' psychometric integrity. The CTT analyses were not used for any decision making, rather they are presented here for the completeness of the psychometric evaluation.

Descriptives

The first step involved the examination of the item descriptive statistics, including M , SD , skewness, and kurtosis. These descriptives, also known as measures of central tendency, help identify poor performing or well performing items. “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” (Kline, 2005, p. 95). Recall that the items were scored dichotomously and therefore the items' means 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 easier items (Tabachnick et al., 2007). Positive skewness indicates that the majority of the scores are clustered around the lower end of the distribution and therefore mark more difficult items (Tabachnick et al., 2007). The skewness of 0 indicates a normal distribution. Skewness at or below -3 and at or above 3 is considered extreme (Kline, 2011). 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, and some do not report kurtosis. With the larger sample sizes such as in this study, the impact of negative or positive kurtosis is diminished (Tabachnick et al., 2007).

Difficulty

Item difficulty was assessed by observing p -values. The p -values represent the percentage of examinees who endorsed the correct response option, and as mentioned above, also represent an item's mean (Kline, 2005). High p -values indicate easy items and low p -values indicate difficult items (Kline, 2005). In general, items that have difficulty of 0 or 1 are useless, because they do not provide variability (information) about examinees. More specifically, items with p -values at or below .20 for all subtests except IC and at or below .25 for IC are considered too difficult and represent a floor effect. Items with p -values at or above .80 for all subtests except IC and at or above .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.

Item Discriminability

Item discriminability, expressed as an item-total correlation (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. None of the items across the cognitive subtests had negative ITC's. With that said, it is important to remember that the new items will be administered with potentially different operational items placed in potentially different places within a subtest. Therefore, we caution against placing much emphasis on the interpretation of these statistics.

Internal Consistency if Item is Deleted

Internal consistency at the subtest level, expressed as Cronbach's alpha, concerns the interrelatedness of items within a subtest (Schmitt, 1996). While the subtest-level alphas are discussed in greater detail elsewhere (see Walsh et al., 2022), the current report provides subtest-level alphas for reference. Internal consistency at the item level shows how the alpha would change if an item was deleted (Schmitt, 1996). Items that contribute to an upward increase in Cronbach's alpha may need to be revised or deleted from the subtest. Generally, alphas below .70 are considered low, between .70-.80 acceptable, and above .80 high (some might argue these alphas might indicate redundancy in measurement such that a reduction of items may improve the measure; Cronbach, 1951; Schmitt, 1996). With that said and similar to item discriminability, it is important to remember that the new items will be administered with potentially different operational items placed in potentially different places within a subtest. Therefore, we caution against placing much emphasis on the interpretation of these statistics.

Item Mean Score Subgroup Differences

For each item, we examined item mean score differences (SGD) between the majority subgroups and legally protected minority subgroups. These differences are estimated using Cohen's d , which is a measure of mean score differences among the two groups expressed in SD units, also known as an effect size (Cohen, 1992). Comparison groups were as follows: Female and Male (F/M), Non-Hispanic Black vs Non-Hispanic White (NHB/NHW), Non-Hispanic Asian vs Non-Hispanic White (NHA/NHW), Non-Hispanic American Indian/Alaska Native vs Non-Hispanic White (NHAIAN/NHW), Non-Hispanic Native Hawaiian/other Pacific Islander vs Non-Hispanic White

(NHNHPI/NHW), Hispanic White vs Non-Hispanic White (HW/NHW), Hispanic Black vs Hispanic White (HB/HW), Multiracial vs Non-Hispanic White (Multi/NHW), and Non-White Hispanic vs. Non-Hispanic White (NWH/NHW).

Conventionally, the magnitudes of .20 would be considered small; .50 would be considered moderate; and .80 or greater would be considered large (Cohen, 1992). The USAF traditionally has interpreted the magnitudes slightly differently (Shore et al., 2020). Specifically, “[T]he Air Force defines the threshold of concern for adverse impact [1] as when subgroup difference reach $d = .40$ ” (Shore et al., 2020, p. 10). “Any discussion of the existence of adverse impact for a measure refers to whether the effect size crosses the threshold of $d = 0.40$ ” (Shore et al., 2020, p. 44). As a compromise between the usual conventions and the most recent studies in the USAF, this report treated Cohen’s d of $\leq .40$ as small (likely not leading to adverse impact); Cohen’s d between .40 and .80 as moderate (likely leading to adverse impact); and Cohen’s $d \geq .80$ as large (highly likely leading to adverse impact). Although these cutoffs were slightly different than the ones prescribed by the literature, they are in line with the cutoffs used for the ASVAB and other standardized assessments throughout the DoD. Moderate-to-large effect sizes suggest that an item is biased against one group over the other and may contribute to adverse impact.

SGD Decision Making

To have confidence in these values, an adequate sampling of both groups was imperative. As such, we only conducted SGD analyses when samples included 10 or more participants from an assessed group. As a result, not all comparison groups were able to be assessed for every sample or subtest. See the “Cohen’s d Sample Sizes” tab for a comprehensive overview of which comparison groups met this cutoff.

Data Sources

The data were collected across three sources – Prolific, Questionmark OnDemand (QM), Officer Training School and Squadron Officer School (OTS/SOS) QM, and OTS/SOS Paper-and-Pencil (P&P). Note that analyses were conducted separately and jointly for each of the data collection resources.

Results

Demographics

Sample demographics depended on the source (e.g., OTS/SOS or Prolific) and on the testing condition (e.g., QM, P&P, QM and P&P). From the racial breakdown perspective, all samples were predominantly White (Form A: 67%-77%; Form B: 64%-76%). From the gender breakdown perspective, there was some variance between samples with OTS/SOS samples being primarily male (Form A: 62%-80%; Form B: 69%-76%) and Prolific samples being primarily female (Form A: 67%-71%; Form B: 54%-58%) except for VA, WK, and BC on Form B. From the ethnic breakdown perspective, all samples were predominantly Non-Hispanic (Form A: 86%-90%; Form

^[1] Adverse impact is defined by the Uniform Guidelines on Employee Selection Procedures as a substantial difference in selection rates between a majority and minority group by a selection test (Equal Employment Opportunity Commission, 1978).

B: 80%-90%). Regarding the socio-economic status, all samples were ‘average’ (Form A: 35%-43%; Form B: 36%-42%). Finally, regarding education levels, OTS/SOS samples had primarily a ‘bachelors’ degree (Form A: 34%-63%; Form B: 55%-57%) and Prolific samples had primarily a ‘high school diploma’ (Form A: 39%-42%; Form B: 37%-40%). See the “Form A Demographics” and the “Form B Demographics” tab for a comprehensive overview of demographic statistics across all subtests and testing conditions.

Descriptives and Difficulty

Average means for the VA new items were generally high, indicating that most participants answered the items correctly (Form A: $M = .80$; Form B: $M = .71$). Nine items on Form A and eight items on Form B showed ceiling effect, suggesting that they were too easy. One item on Form B (VA-16) showed floor effect, suggesting that it was too difficult. All items displayed quite a range of variance, with SD ranging from .11 to .50 between the two forms. In terms of skewness, most items fell within acceptable degrees of skewness, except for the three items on Form A and five items on Form B that exhibited extreme skewness.

Average means for the AR new items were generally high, indicating most participants answered the items correctly (Form A: $M = .81$ Form B: $M = .72$). Ten items on Form A and seven items on Form B displayed a ceiling effect, with no items displaying a floor effect. All items displayed average variance, with SD ranging from .25 to .50 between the two forms. In terms of skewness, most items fell within acceptable degrees of skewness, except for the three items on Form A and one items on Form B that exhibited extreme skewness.

Average means for the WK new items were high, indicating that most participants answered the items correctly (Form A: $M = .82$; Form B: $M = .84$). Twenty items in Form A and twelve items in Form B displayed a ceiling effect, with no items displaying a floor effect. All items displayed a wide range of variance, with SD ranging from .08 to .50 between the two forms. Skewness tended to be negative with twenty items having extreme levels of skewness.

Average means for the MK new items were middling, suggesting a somewhat difficult test (Form A: $M = .57$ Form B: $M = .64$). Only two items on Form B (MK-06 & MK-16) displayed a ceiling effect. No items displayed a floor effect. All items displayed a fairly narrow variance, with SD ranging from .38 to .50 between the two forms. Skewness was generally negative but remained low, with all items staying within acceptable ranges.

Average means for the RC new items were relatively high, suggesting that many participants answered the items correctly (Form A: $M = .68$; Form B: $M = .69$). Eleven items on Form A and eighteen items on Form B displayed a ceiling effect, with no items displaying a floor effect. All items displayed a fairly narrow variance, with SD ranging from .24 to .50 between the two forms. Skewness was generally negative but remained at an acceptable level, except for Form B Item RC-03.

Average means for the PS new items were generally in the middle, suggesting a modestly difficult test (Form A: $M = .51$; Form B: $M = .59$). Three items on Form A and one items on Form B displayed a ceiling effect. Only two items on Form A displayed a floor effect (Item PS-03 and PS-23). All items displayed a fairly wide variance, with SD ranging from .18 to .50 between the two

forms. In terms of skewness, a mix of positive and negative skewness was noticed amongst the items and generally stayed within acceptable levels, except for Form A Item PS-01.

Average means for the TR new items were in the middle, suggesting that some items were endorsed correctly while others were not ($M = .57$). This is not surprising considering that TR is a speeded test. No items exhibited ceiling or floor effects. Again, this is not surprising considering that the new items were placed in the middle of the test (in speeded tests, it is common to observe fairly high p -values at the beginning of a test and fairly low p -values at the end). All items displayed a fairly narrow variance, with SD ranging from .40 to .50. Skewness was generally negative but remained at an acceptable level.

Average means for the IC new items were high, suggesting most participants answered items correctly (Form A: $M = .76$; Form B: $M = .84$). Nine items on Form A and eight items on Form B displayed a ceiling effect. One item on Form A displayed a floor effect (IC-20). All items displayed a fairly wide variance, with SD ranging from .17 to .50 between the two forms. Skewness was generally negative but remained at an acceptable level, except for Form A (IC-03 and IC-10) and Form B (IC-11).

Average means for the BC new items were fairly high, suggesting that the items were relatively easy (Form A: $M = .83$; Form B: $M = .79$). Eleven items on Form A and ten items on Form B displayed a ceiling effect. All items displayed a fairly wide variance, with SD ranging from .13 to .50 between the two forms. In terms of skewness, mostly negative skewness was observed, with some items showing extreme values.

Average means for the AI new items were in the middle, suggesting a modestly difficult test (Form A: $M = .53$; Form B: $M = .48$). Only two items displayed a ceiling effect (Form A AI-05 and AI-14). Three items displayed a floor effect (Form A AI-08, Form A AI-09, and Form B AI-21). All items displayed a fairly wide variance, with SD ranging from .27 to .50 between the two forms. Skewness was generally negative but remained at an acceptable level, except for Form B AI-21.

Discriminability and Internal Consistency

For VA, average ITC's were low to moderate (Form A: $ITC = .41$; Form B: $ITC = .37$). One item on Form B (VA-16) had an ITC of 0. In fact, the same item displayed a floor effect (as mentioned in the previous section). Cronbach's alphas were all in the .80s range which means that the removal of no items resulted in substantial gain or loss of internal consistency. With that said, as mentioned earlier, because the new items will be included along with the existing operational items, the interpretation of these test-dependent and sample-dependent statistics is limited.

For AR, average ITC's were moderate (Form A: $ITC = .47$; Form B: $ITC = .47$). None of the items had negative or zero ITC's. Cronbach's alphas were all in the .80s range which means that the removal of no items resulted in substantial gain or loss of internal consistency.

For WK, average ITC's were low to moderate (Form A: $ITC = .39$; Form B: $ITC = .48$). None of the items had negative or zero ITC's. Cronbach's alphas were all in the .80s range which means that the removal of no items resulted in substantial gain or loss of internal consistency.

For MK, the average ITC's were moderate (Form A: ITC = .46; Form B: ITC = .49). None of the items had negative or zero ITC's. Cronbach's alphas were all in the .80s range which means that the removal of no items resulted in substantial gain or loss of internal consistency.

For RC, average ITC's were low to moderate (Form A: ITC = .36; Form B: ITC = .42). None of the items had negative or zero ITC's, except for one item on Form B (RC-24) with an ITC of .09. Cronbach's alphas were all in the .80s range which means that the removal of no items resulted in substantial gain or loss of internal consistency.

For PS, average ITC's were fairly low (Form A: ITC = .30; Form B: ITC = .33). None of the items had negative or zero ITC's. One item on Form A (PS-01) had an ITC of .07. Notably, Cronbach's alphas were all in the .60s range which means that the items and the subtest itself may have fairly low internal consistency (reliability). However, as described in Section 3.2.5 and shown in Table 9, slight decrements in reliability were expected as a result of maximizing item information for latent ability distribution for the majority subgroup. When new items were incorporated into the two Tv2 parallel forms, Cronbach's alpha for PS was acceptable (.72 and .70, respectively).

For TR, recall that the new items were administered for Form A. The average ITC was high (.79). None of the items had negative or zero ITC's. Cronbach's alphas were all in the .90s range which means that the removal of no items resulted in substantial gain or loss of internal consistency.

For IC, average ITC's were moderate (Form A: ITC = .50; Form B: ITC = .49). One negative ITC was observed for Form A (IC-21). Indications of internal consistency were generally adequate for the items, with the removal of any given new item not resulting in substantial gain to alpha. Cronbach's alphas were all in the .80s range which means that the removal of no items resulted in substantial gain or loss of internal consistency.

For BC, average ITC's for the new BC items, were generally modest (Form A: .49; Form B: .51). Two items displayed ITCs below .30 (Form A: BC-03 & BC-12), suggesting that these items display a poor correlation with the rest of the new BC items. Indications of internal consistency were generally adequate, and each item seemed to not significantly negatively impact alpha, as the removal of any given item did not result in substantial gain to alpha.

Finally, for AI, average ITC's were fairly low (Form A: ITC = .34; Form B: ITC = .41). One negative ITC was observed for Form B (AI-21). Cronbach's alphas for Form A were in the .70s range and for Form B in the .80s range, suggesting that the removal of no items resulted in substantial gain or loss of internal consistency.

SGD

For VA, the sample sizes for six out of seven subgroup comparisons (NB/HW had small sample sizes) were adequate which permitted the SGD analyses. For both forms, most of the items showed small effect sizes across all six comparisons. There were a few items that showed moderate to large effect sizes. Of those, there were two items on Form A and one item on Form B that were consistently flagged across at least two subgroups.

For AR, the sample sizes for six out of seven subgroup comparisons (NB/HW had small sample

sizes) were adequate which permitted the SGD analyses. For both forms, most of the items showed small effect sizes across all six comparisons. There were five items on Form B that were consistently flagged across at least two subgroups.

For WK, the sample sizes for six out of seven subgroup comparisons (NB/HW had small sample sizes) were adequate which permitted the SGD analyses. For both forms, most of the items showed small effect sizes across all six comparisons. There were two items on Form A and one item on Form B that were consistently flagged across at least two subgroups.

For MK, the sample sizes for six out of seven subgroup comparisons (NB/HW had small sample sizes) were adequate which permitted the SGD analyses. For both forms, most of the items showed small effect sizes across all six comparisons. There were two items on Form A and three items on Form B that were consistently flagged across at least two subgroups.

For RC, the sample sizes for six out of seven subgroup comparisons (NB/HW has small sample sizes) were adequate which permitted the SGD analyses. For both forms, most of the items showed small effect sizes across all six comparisons. There was one item on Form A and ten items on Form B that were consistently flagged across at least two subgroups.

For PS, the sample sizes for six out of seven subgroup comparisons (NB/HW had small sample sizes) were adequate for Form A and all seven were adequate for Form B. For both forms, most of the items showed small effect sizes across all comparisons. There was one item on Form B that was consistently flagged across at least two subgroups.

For TR Form A, the sample sizes for six out of seven subgroup comparisons (NB/HW had small sample sizes) were adequate which permitted the SGD analyses. Most of the items showed small effect sizes across all six comparisons. There were a few items that showed moderate effect sizes for Form A, one of which was consistent between two groups (TR-26).

For IC, the sample sizes for six out of seven subgroup comparisons (NB/HW had small sample sizes) were adequate which permitted the SGD analyses. For both forms, most of the items showed moderate to large effect sizes across the comparisons. Compared to the rest of the subtests, IC emerged as most problematic in terms of mean score SGD. There were twelve items on Form A and seven items on Form B that were consistently flagged across at least two subgroups.

For BC, the sample sizes for six out of seven subgroup comparisons (NB/HW had small sample sizes) were adequate which permitted the SGD analyses. For both forms, most of the items showed small effect sizes across all six comparisons. There were a few items that showed moderate effect sizes but none of them were consistent across subgroups.

For AI, the sample sizes for six out of seven subgroup comparisons (NB/HW had small sample sizes) were adequate which permitted the SGD analyses. For both forms, most of the items showed small effect sizes across all six comparisons. There were three items on Form A and eleven items on Form B that were consistently flagged across at least two subgroups.

Discussion

Considering that CTT produces test- and sample-dependent statistics, the results of the various analyses that we present in this appendix and in Attachment 1 “CTT Parameters.xlsx” are to be interpreted with caution. In fact, we recommend that decisions regarding which items to include in Form Tv2 should be based on the IRT analyses described in the body of the technical report and not on the CTT results described here.

With that said, the general pattern of the CTT results suggests that the new items were fairly easy for VA, AR, WK, BC, and IC. The items in the rest of the subtests were either of moderate or higher difficulty, with the highest difficulty observed for Form B AI items. This pattern of results is not surprising given that the first half of the AFOQT targets people’s academic aptitudes and therefore Prolific and OTS/SOS samples should find them fairly easy. The second half of the AFOQT targets people’s specialized knowledge (e.g., aviation, PS), which may indicate motivation. Therefore, it is not surprising that Prolific samples may find these subtests more challenging than would OTS or SOS samples.

In terms of discriminability, most of the items fell within the acceptable range of ITC’s, with a few exceptions when the ITC’s were either null or negative. Again, this could be a function of an unrepresentative sample or of a mixture of items that do not fit together. Using IRT, appropriate items can be included among the operational tests to ensure requisite discriminability levels (*‘a’* parameters).

Finally, in terms of mean score SGD, most items within and between subtests and the two forms showed small effect sizes. Perhaps, the subtests with the highest number of items that were flagged as having either moderate or large effect size among the various comparison subgroups were IC and AI. This is not surprising in light of our speculation earlier that these tests are more specialized and therefore would require some preparation or motivation on the part of the test-takers. Socio-economic status may also play role in the performance on these subtests.

Conclusion

This appendix describes the methods and results of the CTT analyses performed on the new items that were developed as part of the Enhanced Airman Alignment program Task 13 AFOQT Form Tv2 effort to replace the problematic items on the AFOQT Form Tv1. While the CTT results described here are informative, we recommend that decisions regarding which items to include in Form Tv2 should be based on the IRT analyses described in the body of the technical report and not on the CTT results presented here.

References

- Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, 1, 98-101.
- Cronbach, L. H. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, 2, 292-307.

- Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Fed Register*, 1, 216-243.
- Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics and the importance of psychological measurement*. *Psychometrics*. Thousand Oaks, CA: Sage.
- Kline, T. J. B. (2005). *Psychological testing*. Thousand Oaks, CA: Sage.
- Kline, R. B. (2011). *Principles and practices of structural equation modeling*. (4th ed.). NY: Guilford Press.
- Nunnally, J. C. & Bernstein, I. R. (1994). *Psychometric theory*. (3rd ed.). NY: McGraw-Hill.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8, 350-353.
- 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.
- 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: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
- 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. Vandenberg (Eds.). *Statistical and methodological myths and urban legends, Doctrine, verity and fable in the organizational and social sciences* (pp. 37-59). NY: Routledge/Taylor & Francis.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

χ^2	Chi-Square
r	Correlation
ABM	Air Battle Manager
ABM_Max-Old	Difference between Maximum of Form T ABM and Form Tv2 ABM Composite Score and Form T ABM Score
ABM_New-Old	Difference between Form Tv2 ABM and Form T ABM Composite Score
AFOQT	Air Force Officer Qualifying Test
AFPC/DSYX	Air Force Personnel Center Strategic Research and Analysis Branch
AFPC	Air Force Personnel Center
AI	Aviation Information
APAT	Applied Performance Assessment and Testing
AR	Arithmetic Reasoning
ASVAB	Armed Services Vocational Aptitude Battery
BC	Block Counting
BMT	Basic Military Training
CSO	Combat Systems Operator
CSO_Max-Old	Difference between Maximum of Form T CSO and Form Tv2 CSO Composite Score and Form T Score
CSO_New-Old	Difference between Form Tv2 CSO and Form T CSO Composite Score
CTT	Classical Test Theory
D&I	Diversity and Inclusion
DI	Diversity Index
DIF	Differential Item Functioning
EM	Expectation Maximization
Form Tv1	AFOQT Form T, version 1
Form Tv2	AFOQT Form T, version 2
IC	Instrument Comprehension
IFT	Initial flight Training
IIF	item information functions
IRT	Item Response Theory
MH	Mantel-Haenszel
M	Mean
MASS	Merit Assignment Selection System
MK	Math Knowledge
NR	Number Right
OTS	Officer Training School
Pilot_Max-Old	Difference between Maximum of Form T Pilot and Form Tv2 Pilot Composite Score and Form T Score
Pilot_New-Old	Difference between Form Tv2 Pilot and Form T Pilot Composite Score
P&P	Paper-and-Pencil
PS	Physical Science
RC	Reading Comprehension
RiU	Reduction in Uncertainty

SD	Standard Deviation
SDI-O	Self-Description Inventory
SGD	Subgroup Difference
SJT	Situational Judgment Test
SOS	Squadron Officer School
SUPT	Supervised Undergraduate Pilot Training
TR	Table Reading
USAF	United States Air Force
VA	Verbal Analogies
WK	Word Knowledge