



AFRL-RH-WP-TR-2022-0022

**AIR FORCE OFFICER QUALIFYING TEST (AFOQT)
FORM T EVALUATION:
SUBTEST-LEVEL ANALYSES**

**Julia L. Walsh
Montana R. Woolley
Michael F. Brady
Infoscitex, a DCS company**

**Thomas R. Carretta
711 HPW/RHBC**

**March 2022
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-2022-0022 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

CARRETTA.T
HOMAS.R.12
28984929

Digitally signed by
CARRETTA.THOMAS.R
.1228984929
Date: 2022.05.02
12:51:00 -04'00'

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

MCKINLEY.RICH
ARD.A.12622397
10

Digitally signed by
MCKINLEY.RICHARD.A.1262
239710
Date: 2022.05.05 13:03:42
-04'00'

R. ANDY MCKINLEY, DR-III, PhD
Core Research Area Lead,
Cognitive and Physical Performance
Performance Optimization Branch
Airman Biosciences Division

This report is published in the interest of scientific and technical information exchange, 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) 31-03-22		2. REPORT TYPE Interim		3. DATES COVERED (From - To) 14 January 2021 – 28 February 2021	
4. TITLE AND SUBTITLE Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: Subtest-Level Analyses				5a. CONTRACT NUMBER FA8650-20-F-4094	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Julia L. Walsh ^a , Montana, R. Woolley ^a , Michael F. Brady ^a , Thomas R. Carretta ^b				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER H12E	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) ^a Infoscitex, a DCS Company 4027 Colonel Glenn Highway, Suite 210 Dayton, OH 45431-1672				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) ^b Air Force Materiel Command Air Force Research Laboratory 711 th Human Performance Wing Airman Systems Directorate Airman Biosciences Division Performance Optimization Branch Wright-Patterson AFB, OH 45433				10. SPONSORING/MONITORING AGENCY ACRONYM(S) 711 HPW/RHBC	
				11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S) AFRL-RH-WP-TR-2022-0022	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution Statement A: Approved for public release.					
13. SUPPLEMENTARY NOTES Report contains color. AFRL-2022-1585, cleared 26 April 2022					
14. ABSTRACT The Air Force Officer Qualifying Test (AFOQT) is used to qualify applicants for officer commissioning and for aircrew training as pilots, combat system officers, air battle managers, and remotely-piloted aircraft pilots. The purpose of this report is to summarize psychometric evaluation of the AFOQT Form T at the subtest level. Data consisted of responses from United States Air Force officer applicants who were administered either AFOQT Form T1 or Form T2 between 2014 and 2020. Overall, the subtest-level results were consistent with previous research in terms of psychometric properties (Morath et al., 2008; Carretta et al., 2016; Aguilar, 2017). However, there were areas where improvements could be made. For example, subtest-level analyses revealed that Reading Comprehension and Table Reading are the easiest subtests with low discriminability. Moderate to large mean score subgroup differences were also observed. Results of the confirmatory factor analyses revealed a better model-data fit for a bifactor model with five content factors representing verbal, math, spatial, perceptual speed, and aviation knowledge, and a general factor. Recommendations for improvement are provided.					
15. KEY WORDS Air Force Officer Qualifying Test, AFOQT, subtest-level analyses					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT: SAR	18. NUMBER OF PAGES 122	19a. NAME OF RESPONSIBLE PERSON (Monitor) Thomas R. Carretta
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (Include Area Code) N/A

Table of Contents

EXECUTIVE SUMMARY	iv
1.0 INTRODUCTION	1
1.1 Brief History of the AFOQT	1
1.1.1 AFOQT Form T	1
1.2 Initial and Current Psychometric Evaluations	2
1.2.1 Morath et al. (2008)	2
1.2.2 Carretta et al. (2016)	3
1.2.3 Aguilar (2017)	3
1.2.4 The Present Research	3
2.0 METHOD	3
2.1 Sample Characteristics	4
2.2 Technical Attachments	6
2.3 Scoring	6
2.4 Data Scrubbing	6
2.4.1 Missingness	6
2.4.2 Carelessness	7
2.5 Measures	8
2.6 Technical Approach	8
2.7 Organization of the Report	10
3.0 RESULTS	10
3.1 RQ1: What Are the Psychometric Properties of the Subtests?	10
3.1.1 Descriptives	10
3.1.2 Difficulty	11
3.1.3 Percentiles	12
3.1.4 Speededness and Overt Guessing	13
3.1.5 Test Information Function	15
3.1.6 Reading Grade Level	18
3.1.7 Discriminability	20
3.2 RQ2: How Do the Subtests Relate to Each Other?	20
3.2.1 Correlations	21
3.2.2 Confirmatory Factor Analysis	22
3.3 RQ3: What Are the Mean Score Subgroup Differences?	26
3.3.1 Cohen's <i>d</i>	26
3.3.2 Measurement Invariance	32
3.3.3 Variance Explained by Demographic Variables	34
3.4 RQ4: Are the Subtests Reliable?	36
3.4.1 Internal Consistency	36
3.4.2 Test-Retest Reliability	37
3.5 RQ5: Are T1 and T2 Equivalent?	39
3.5.1 Methodology	39
3.5.2 Between-Subjects Alternate Forms Reliability Results	40
3.5.3 Within-Subjects Alternate Forms Reliability Results	41
3.5.4 LMEM Results	42
3.6 RQ6: Who Retests? Do Retesters Perform Better When They Retest?	43

3.6.1	First Attempt Scores for Retesters vs. Non-Retesters.....	43
3.6.2	Demographic Differences between Retesters and Non-Retesters	44
3.6.3	Effects of Retesting on Scores	46
3.6.4	Score Improvements between Multiple Retests.....	48
3.7	RQ7: Was There a Test Security Breach?	50
4.0	DISCUSSION	52
4.1	Summary of Psychometric Properties.....	53
4.2	Summary of Subtest Inter-Correlations	53
4.3	Summary of Mean Score Subgroup Differences	53
4.4	Summary of Reliability	54
4.5	Summary of Equivalency.....	54
4.6	Summary of Retesting.....	54
4.7	Summary of Test Security	54
5.0	CONCLUSION.....	55
6.0	REFERENCES	56
	APPENDIX A: Stratified Samples Demographics	60
	APPENDIX B. Statistics and Their Interpretation	61
	APPENDIX C. Stratified Samples Descriptives.....	62
	APPENDIX D. Stratified Samples Difficulty.....	64
	APPENDIX E. Stratified Samples Percentiles	65
	APPENDIX F. Stratified Samples Speededness.....	68
	APPENDIX G. Stratified Samples Test Information Functions.....	72
	APPENDIX H. Stratified Samples Discriminability	74
	APPENDIX I. Stratified Samples Correlations	75
	APPENDIX J. CFA Path Diagrams.....	78
	APPENDIX K. Stratified Samples Subgroup Differences	80
	APPENDIX L. Measurement Invariance Chi Square Tests	94
	APPENDIX M. Stratified Samples Internal Consistency.....	95
	APPENDIX N. Stratified Test-Retest Reliabilities	96
	APPENDIX O. Stratified Samples Between-Subj. Alt. Forms Reliability – All Items.....	98
	APPENDIX P. Stratified Samples Between-Subj. Alt. Forms Reliability – Common Items	100
	APPENDIX Q. Stratified Samples Within-Subj. Alt. Forms Reliability – All Items	102
	APPENDIX R. Stratified Samples Retesting Effects on the Same Form – T1T1	104
	APPENDIX S. Stratified Samples Retesting Effects on the Same Form – T2T2	106
	APPENDIX T. Stratified Sample Retesting Effects on the Parallel Forms.....	108
	APPENDIX U. Stratified Samples Stability Analysis.....	110
	LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS.....	113

LIST OF FIGURES

Figure 1. Flesch-Kincaid Formula	18
--	----

LIST OF TABLES

Table 1. AFOQT Form T Subtests.....	1
Table 2. Unstratified Sample Demographics	5
Table 3. Research Questions and Associated Analyses.....	9
Table 4. Unstratified Sample Descriptives	10
Table 5. Unstratified Sample Difficulty – p-values	12
Table 6. Unstratified Sample Percentiles.....	13
Table 7. Unstratified Sample Speededness	14
Table 8. Unstratified Sample Overt Guessing	15
Table 9. Unstratified TIF	16
Table 10. Unstratified Sample RC RGL Results	19
Table 11. Unstratified Sample Discriminability - ITC	20
Table 12. Unstratified Sample Bivariate Correlations.....	22
Table 13. Unstratified Sample CFA Model-Data Fit Indices	25
Table 14. Unstratified Sample Mean Score Subgroup Differences (Cohen’s d).....	28
Table 15. MI Model-Data Fit Indices	33
Table 16. Unstratified Sample Stepwise Regression Results	35
Table 17. Unstratified Sample Internal Consistency (Cronbach’s Alpha)	36
Table 18. Unstratified Sample Test-Retest Reliability	38
Table 19. Unstratified Sample Between-Subj. Alt. Forms Reliability – All Items	40
Table 20. Unstratified Sample Between-Subj. Alt. Forms Reliability – Common Items.....	41
Table 21. Unstratified Sample Within-Subj. Alt. Forms Reliability – All Items	41
Table 22. Unstratified Sample Alt. Forms Reliability – LMEM	43
Table 23. Unstratified Sample First Attempt Scores for Non-Retesters vs. Retesters	44
Table 24. Unstratified Sample Non-Retesters vs. Retesters: Age and Education	44
Table 25. Unstratified Sample Non-Retesters vs. Retesters: Demographic Frequencies.....	45
Table 26. Unstratified Sample Retesting Effects on the Same Form	47
Table 27. Unstratified Sample Retesting Effects on the Parallel Forms	48
Table 28. Unstratified Sample Two and Three Attempt Score Increases.....	49
Table 29. Unstratified Sample Stability Analysis.....	51
Table 30. Subtest Summary	52

EXECUTIVE SUMMARY

The purpose of the analyses summarized in this technical report was to psychometrically evaluate the 10 cognitive subtests included in the Air Force Officer Qualifying Test (AFOQT) Form T. Below is a brief summary of the findings and recommendations moving forward.

First, the results for the Form T cognitive subtests were consistent with previous research in terms of measures of central tendency, difficulty, and discriminability – both within and between the parallel forms (T1 and T2; Morath et al., 2008; Carretta et al., 2016; Aguilar, 2017). Second, all Form T cognitive subtests related to one another and were g-loaded. Third, consistent with the broader literature on cognitive ability, mean score differences were observed for Form T across the gender, race, and ethnic subgroups. The largest effect sizes were observed for Black test-takers, followed by female test-takers, and then by Asian test-takers, compared to their racial and gender majority subgroup counterparts. Fourth, all subtests on Form T showed moderate-to-high internal consistency and test-retest reliabilities. Fifth, all subtests showed acceptable alternate forms reliabilities, with sufficient evidence for the equivalency of the parallel forms, rendering the issue of equating unnecessary. Sixth, when comparing retesters and non-retesters, the former tended to score lower on their first attempt, which subsequently led them to retake the test. Retesters' scores improved with each attempt. Finally, because test security is always a concern, stability analyses were performed. The conclusion was that a breach in test security was unlikely to have occurred.

Based on these findings, we recommend the following: (1) the cognitive subtests should be designed to target specific ranges of cognitive ability (preferably, around cutoff scores) and with sufficient discriminability to separate lower and higher ability test-takers; (2) each subtest should measure its respective officership-critical competency to avoid redundancy; (3) the AFOQT should include both cognitive (preferably, a mix of crystallized and fluid intelligence) and non-cognitive measures that have been shown to reduce subgroup differences, such as personality and situational judgment tests; (4) the subtests should include an optimal number of items to ensure measurement precision but reduce redundancy; (5) the subtest scores should be equated if sufficient evidence of non-equivalency between the parallel forms is found (note that equating was not necessary for this study, but may be necessary for future studies); (6) testing policy should allow for retesting; and (7) test security should be continuously evaluated.

While these results and recommendations are meant to inform decisions regarding future iterations of the AFOQT, we caution that the ultimate decision about the inclusion, exclusion, or revision of the current test needs to be based on the holistic interpretation of the data resulting from: Item-level analyses (Walsh et al., in press), composite-level analyses, criterion-related validities, adverse impact¹ rates and subgroup qualification rates, critical officer job analyses (Kantrowitz et al., in press), and analyses for non-cognitive subtests, including the Situational Judgment Test (Walsh et al., 2022) and the Self-Description Inventory (Woolley et al., in progress).

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).

1.0 INTRODUCTION

This technical report describes the results of the subtest-level analyses performed to evaluate the psychometric properties of the 10 cognitive subtests in the AFOQT Form T. This introductory section briefly overviews the history of the AFOQT, describes the composition of the Form T, and lays out the purpose of the present research.

1.1 Brief History of the AFOQT

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

1.1.1 AFOQT Form T

The current AFOQT Form T became operational on August 1, 2014. There are two parallel forms of Form T (identified as T1 and T2) to reduce practice effects and to preserve test security. It has 10 cognitive and two non-cognitive subtests and takes approximately 4 hours to administer. Table 1 summarizes the AFOQT Form T subtests with their administration time and number of items.

Table 1. AFOQT Form T Subtests

Subtest (in the order of administration)	Subtest Abbreviation	Subtest Type	Subtest Length (in min)	Number of Items
1. Verbal Analogies	VA	Cognitive	8	25
2. Arithmetic Reasoning	AR	Cognitive	29	25
3. Word Knowledge	WK	Cognitive	5	25
4. Math Knowledge	MK	Cognitive	22	25
5. Reading Comprehension	RC	Cognitive	38	25
6. Situational Judgment Test	SJT	Non-Cognitive	35	50
7. Self-Description Inventory - Officer	SDI-O	Non-Cognitive	45	240
8. Physical Science	PS	Cognitive	10	20
9. Table Reading	TR	Cognitive	7	40
10. Instrument Comprehension	IC	Cognitive	5	25
11. Block Counting	BC	Cognitive	4.5	30
12. Aviation Information	AI	Cognitive	8	20

Subtest scores are combined to generate composite scores used to predict success in certain types of United States Air Force (USAF) training programs. Note that the PS, SJT, and SDI-O subtests are not currently included in the operational composites. Nine of the 10 cognitive subtests make up six operational composites as shown below:

1. Pilot: MK, TR, IC, AI
2. CSO: WK, MK, TR, BC
3. ABM: VA, MK, TR, IC, BC, AI
4. Academic Aptitude: VA, AR, WK, MK, RC
5. Verbal: VA, WK, RC
6. Quantitative: AR, MK

USAF cadets, enlisted personnel, and civilians from a variety of accession sources including Officer Training School-Civilian (OTS-CIV), Officer Training School-Active Duty (OTS-AD), AECP, United State Air Force Academy (USAF A), Reserve Officer Training Corps (ROTC), Air National Guard (ANG), and Air Force Reserves (AFRES). All these sources, except the USAFA, take the AFOQT to apply for officer commissioning. All test-takers regardless of commissioning source use the AFOQT scores to qualify for classification into a rated career field, including CSO, ABM, fighter or mobility pilot, and Remotely Piloted aircraft (RPA) pilot. Each applicant currently has an opportunity to take the assessment up to two times (i.e., the first attempt and a re-take) with an opportunity to take the assessment a third time if a request for a waiver is granted. Test-takers were historically required to wait 180 days between attempts. This policy was subsequently amended to 150 days (Fedrigo, 2021) and has recently been reduced to 90 days.

1.2 Initial and Current Psychometric Evaluations

The AFOQT undergoes psychometric evaluation on a periodic basis. Form S was evaluated by Morath et al. (2008) and Form T by Carretta et al. (2016). Aguilar (2017) subsequently compared Forms S and T. The present report draws parallels among the findings reported by these reports, where applicable and available.

1.2.1 Morath et al. (2008)

Data for the Morath et al. (2008) effort consisted of responses from USAF officer applicants who were administered either the AFOQT Form S1 ($N = 9,607$) or Form S2 ($N = 8,164$) between 2005 and 2007. In general, both forms demonstrated acceptable subtest-level psychometric properties in terms of internal consistency, test-retest reliability, stability analyses, and alternate forms reliability. There were subgroup differences observed among the gender, ethnic, and racial minority test-takers and majority test-takers between and within the parallel forms. Note that there were some qualitative changes between Form S and Form T. Therefore, the comparisons we will provide are to be interpreted with caution. Specifically, Form S included two spatial subtests (Rotated Blocks and Hidden Figures) that are not included in Form T. Reading Comprehension (RC) was not administered in Form S and was added to Form T. General Science (GS) in Form S was modified to focus on Physical Science (PS) on Form T. For IC, the number of items increased from 20 to 25, and testing time was shortened from 6 to 5 minutes; aircraft and instrument graphics were modernized and subtest instructions were updated for clarity. For BC, the number of items increased from 20 to 30, testing time increased proportionately from 3 to 4.5 minutes, and

instructions were updated for clarity (see Aguilar, 2017) for more details on the differences between Form S and Form T.

1.2.2 Carretta et al. (2016)

Data for the Carretta et al. (2016) effort consisted of responses from USAF officer applicants who were administered either the AFOQT Form T1 ($N = 5,681$) or Form T2 ($N = 5,199$) between 2015 and 2016. In general, both forms demonstrated acceptable subtest-level psychometric properties. However, there were a few problematic subtests. For example, there were a few items on BC with the blocks touching on their back sides out of the participant's view. None of the example items illustrated this condition, thereby artificially making the subtest more difficult. Results of the confirmatory factor analyses were consistent with previous forms, such that a model with five-factors representing verbal, math, spatial, perceptual speed, and aviation knowledge and a hierarchical general factor showed the best fit. However, some fit indices were below desired levels which could have been caused by the non-normality of the subtest scores and the underrepresentation of the spatial factor (Carretta et al., 2016).

1.2.3 Aguilar (2017)

Using the same dataset as in Carretta et al. (2016), Aguilar (2017) performed psychometric analyses comparing Form T (T1/T2) and Form S (S1/S2). The properties of Form T and Form S were generally similar, with no significant changes in reliability between the two forms. Means and standard deviations showed VA, AR, WK, and MK had a higher difficulty on Form T than on Form S. Skewness values indicated a reduction in ceiling effects for various subtests on Forms T1/T2 as compared to Forms S1/S2 (Aguilar, 2017).

1.2.4 The Present Research

The present research was initiated by the Air Force Personnel Center Strategic Research and Assessments Branch (AFPC/DSYX), who after collecting more operational data was interested in evaluating the psychometric integrity of the AFOQT Form T cognitive and non-cognitive subtests at the item-, subtest-, and composite-levels and to make recommendations for potential changes and enhancements to future revisions of the test. Details of the item-level (Walsh et al., in press) and composite-level analyses (Kantrowitz et al., in press) for the cognitive subtests and of the item-level and subtest-level, analyses for the SJT (Walsh et al., 2022) and the SDI-O (Woolley et al., in progress) are presented in several different technical reports. This report serves as the final deliverable for the subtest-level analyses performed on the 10 cognitive subtests.

2.0 METHOD

Analyses were performed using AFOQT Form T (T1/T2) data collected between 2014 and 2020 from the test-takers who either were applying for USAF officer commissioning or for classification into a rated officer career field (ABM, CSO, Pilot, and RPA Pilot).

2.1 Sample Characteristics

Table 2 shows sample sizes (N) and proportions (%) of each demographic category for Forms T1 and T2. The total sample size for T1 was 36,099 and for T2 was 32,952. Note that when a demographic is reported as ‘Unknown,’ it indicates that a test-taker either declined to respond or that their response was not legibly marked (i.e., did not select a response). The demographic composition of the two samples was similar. The mean age of test-takers was 23.79 ($SD = 5.04$) for T1 and was 23.65 ($SD = 5.00$) for T2. The mean education levels were 14.88 years for T1 and 14.82 years for T2. Both samples predominantly consisted of male (74% for T1 and T2) and White (77% for T1 and T2) identifying test-takers. Appendix A contains the demographic information stratified by accession source. The differences across the accession sources in terms of demographics are described in more detail in Section 3.3.

Table 2. Unstratified Sample Demographics

Sample Characteristic	T1		T2	
	N	%	N	%
Sex				
Male	26,566	74%	24,286	74%
Female	9,497	26%	8,618	26%
Unknown	36	0%	48	0%
Race*				
American Indian/Alaska Native	1,939	5%	1,812	5%
Asian	3,729	10%	3,310	10%
Black/African American	5,005	14%	4,577	14%
Native Hawaiian/Other Pacific Islander	918	3%	825	3%
White	27,849	77%	25,255	77%
Multiracial	3,982	11%	3,616	11%
Ethnicity				
Hispanic	5,212	14%	4,745	14%
Non-Hispanic	30,211	84%	27,505	83%
Unknown	676	2%	702	2%
Socio-Economic Status				
Much higher than average	1,751	5%	1,706	5%
Somewhat higher than average	10,197	28%	9,421	29%
Average	14,079	39%	12,543	38%
Somewhat lower than average	6,216	17%	5,594	17%
Much lower than average	2,113	6%	2,009	6%
Declined to respond	1,629	5%	1,576	5%
Unknown	114	0%	103	0%
Years of Education Completed				
12	1,177	3%	1,251	4%
13	10,382	29%	9,994	30%
14	4,847	13%	4,314	13%
15	3,863	11%	3,368	10%
16	10,790	30%	9,441	29%
17	2,294	6%	2,068	6%
18	1,849	5%	1,697	5%
19	399	1%	359	1%
20	218	1%	204	1%
21+	221	1%	190	1%
Unknown	59	0%	66	0%
Highest Academic Degree				
High School Diploma	17,420	48%	16,359	50%
Associate's Degree	3,741	10%	3,393	10%
Bachelor's Degree	13,023	36%	11,402	35%
Master's Degree	1,692	5%	1,628	5%
Doctoral Degree	128	0%	89	0%
Unknown	95	0%	81	0%
Accession Source				
OTS-CIV	8,026	22%	7,284	22%
OTS-AD	5,430	15%	4,843	15%
AACP	170	0%	168	1%
USAFA	4,312	12%	3,741	11%
ROTC	12,765	35%	12,413	38%
ANG	3,915	11%	3,288	10%
AFRES	1,481	4%	1,215	4%

*The proportions for Race do not add to 100% because test-takers had an option to choose more than one race. Those who chose two or more races are summarized by the multiracial category.

Note. N = Sample Size.

2.2 Technical Attachments

Note that this technical report refers to three technical attachments. Attachment 1 contains a detailed description of the data scrubbing steps, summarized in Sections 2.4.1 and 2.4.2. Attachment 2 contains complete output for T1 and T2 for stratified and unstratified samples. Attachment 3 contains a short overview of the conventional cutoffs for Cohen's d effect sizes and the cutoffs adopted in the current research (see Section 3.3).

2.3 Scoring

Subtest scores were created by applying a scoring key provided by AFPC/DSYX (0 = incorrect or missing; 1 = correct) to the item-level data and averaging the items within each subtest.

2.4 Data Scrubbing

The AFPC/DSYX performed the initial data scrubbing. Specifically, cases were deleted listwise if: (1) they had an invalid social security number; (2) the test was scanned and recorded more than once; (3) the date of testing was indecipherable or incomplete; or (4) the test administrator flagged the test as problematic. *Listwise deletion* means that the cases are deleted altogether and are not included in any analyses (Allison, 2001). Next, the data were de-identified (i.e., personally identifiable information [PII] was removed). Each test-taker was assigned a unique encrypted alpha-numeric identifier (i.e. SPARKIDs).

The initially scrubbed dataset that IST received from the AFPC/DSYX included 78,397 cases across T1 and T2. Note that several items were discovered to have been mis-keyed – T1 PS Items 13, 16, and 17; T2 PS Item 14; and T2 RC Item 12. The items were re-keyed and the subtest scores for the impacted subtests were re-calculated. IST removed listwise the test-takers who self-identified as belonging to the Test Control Officer-Test Evaluator (TCO-TE), Active Duty Other Accession Source (AD-Other), or Civilian Other Accession Source (CIV-Other) accession sources, because they did not constitute the target population. If the accession source was not legibly marked (i.e., asterisk or blank), cases were removed listwise due to inability to categorize them into the appropriate stratified sample group. Next, the variables irrelevant to the present research were removed to keep only the metadata, the demographics data, and the item and subtest scores. Finally, unit missingness and unit careless responding were removed listwise. Sections 2.4.1 and 2.4.2 describe how we handled missingness and careless responding in more detail.

2.4.1 Missingness

Data missingness is a common and pervasive problem in behavioral and social sciences. It adds noise to valid data and may result in biased statistical conclusion validity (Cook et al., 2002). Therefore, it was important to examine the nature of missingness in the AFOQT data.

Missingness is classified as unit nonresponse or item nonresponse. *Unit-nonresponse* occurs when the information is available for the unit but not for the variables (e.g., a test-taker provided their demographic data, but not the responses on a subtest; Schafer & Graham, 2002). *Item-nonresponse* occurs when the information is available for some items but not for other items (e.g., a test-taker

provided responses to 10 of 25 items on the SJT; Schafer & Graham, 2002).

Missingness may be due to random and nonrandom factors. *Random factors* include forgetting to respond to an item or a response to an item that is illegibly marked for paper-and-pencil tests such as the AFOQT Form T (Huisman, 1999; Schafer & Graham, 2002). *Nonrandom factors* include cheating/faking, fatigue, item dependency, and motivation (Huisman, 1999; Schafer & Graham, 2002).

There are three common terms associated with data missingness – Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) (Schafer & Graham, 2002). *MCAR* means that a missing response on variable X for a test-taker does not depend on their own values on Variable X and another Variable Y (and, by independence, does not depend on Variable X and Y of other test-takers). *MAR* means that the missing response on Variable X may or may not depend on Variable Y, but not on Variable X. *MNAR* means that the missing response on Variable X depends on Variable X. MNAR is least preferred of the three missingness types and MCAR is most preferred.

The examination of missingness at the subtest-level revealed that the percentage of missing data increased throughout most of the subtests (i.e., a lower percentage of missing data in the first half of the subtest and a higher percentage of missing data in the second half of the subtest). Three subtests – TR, BC, and IC – showed missingness at a slightly higher rate than a common rule of thumb of five percent for data MCAR (Schafer & Graham, 2002). However, because these subtests were designed to be speeded, missingness simply indicates that the test-takers ran out of time.

2.4.2 Carelessness

Test-taking motivation is defined as “willingness to engage as well as invest effort and persistence in working on test items” (Rios et al., 2017, p. 74). Low test-taking motivation may result in unit-nonresponse (as described above) or in responses which do not reflect the test-taker’s best effort. The latter has been termed either *careless responding* or *insufficient effort responding* (IER; Huang et al., 2012). Potential impacts of IER include lower validity of score-based inferences, biased individual ability estimates, biased item parameter estimates, biased reliability estimates, and biased correlations with external variables (Huang et al., 2012).

Many methods can be used to detect IER, each of which can capture different types of IER with different levels of sensitivity (Huang et al., 2012; Meade & Craig, 2012). These methods include pre-test and post-test strategies. *Pre-test strategies* include instructed response items, infrequency items, and self-reported items (Meade & Craig, 2012). *Post-test strategies* can be organized into response-pattern-based methods, response-latency-based methods, and individual-reliability-based methods (Meade & Craig, 2012).

Since the present research utilized archival data, only post-test strategies were considered. Of the three aforementioned post-test strategies, the response-latency-based method was not possible because the AFOQT Form T is administered as a paper-and-pencil test. Thus, the other two methods were considered. Ultimately, we decided to only use unit longstring – a response-pattern-based method that identifies straight-line response patterns throughout the subtest (e.g., responding

with “C” for all items in a test). This method is simple and captures the ‘low-hanging fruit’ (i.e. the worst and most obvious type of carelessness). We used this maximum level of longstring in order to err on the side of Type II errors, and avoid removing potentially valid responses from the dataset.

After applying the aforementioned data scrubbing steps, and removing unit careless responding and unit missingness, the initial dataset was split into several datasets to accommodate the various specialized analyses (subgroup differences, test-retest reliability, alternate forms reliability, etc.). Note that these steps resulted in the subtests having slightly different sample sizes across the analyses and across stratifications. The present report includes only the overall sample sizes for the stratified and unstratified samples for each analysis, not broken down by each subtest. For a step-by-step data scrubbing process and for the subtest-specific sample sizes, please refer to the Attachment 1 attached to the report.

2.5 Measures

Below is a summary of the constructs that the 10 cognitive subtests were designed to measure (Carretta et al., 2016):

- VA assesses the ability to reason and determine the relations between words;
- AR uses word problems to assess the ability to understand arithmetic relations;
- WK measures verbal comprehension of written language involving the use of synonyms;
- MK assesses the ability to use mathematical formulas, relations, and terms;
- RC assesses the ability to read and understand written material;
- PS provides a measure of knowledge and understanding of scientific, terms, concepts, instruments, and principles;
- TR measures the ability to quickly and accurately extract information from tables;
- IC measures the ability to determine the altitude of an aircraft from illustrations of flight instruments;
- BC provides a measure of spatial ability through the analysis of three-dimensional representation of a set of blocks;
- AI assesses knowledge of general aviation concepts, principles, and terms.

2.6 Technical Approach

The analyses were restricted to first-time test-takers, unless otherwise indicated. Note that the analyses were performed on both unstratified and stratified samples. The unstratified sample includes all data; whereas the stratified samples include data for one of the seven accession sources: (1) OTS-CIV, (2) OTS-AD, (3) AECF, (4) USAFA, (5) ROTC, (6) ANG, and (7) AFRES. Note that the results presented throughout this section are for the unstratified sample, unless otherwise stated. The results for the stratified samples are included in the appendices.

Although the performance work statement did not specify any particular hypotheses, IST identified a series of research questions (RQ’s) to evaluate the psychometric integrity of the AFOQT Form T subtests. The approach used to examine AFOQT subtest effectiveness was grounded in Classical Test Theory (CTT) and in Item Response Theory (IRT). Table 3 provides a brief description of

the analyses that were planned for addressing each RQ. Appendix B contains a summary of the statistics and their interpretation.

Table 3. Research Questions and Associated Analyses

RQ's	Section	Analyses
What are the psychometric properties of the subtests?	3.1	For each subtest, the following statistics were generated: The CTT measures of central tendency (means, standard deviations, skewness, and kurtosis); the CTT difficulty statistics (p -values); the percentiles; the IRT Test Information Function; the CTT discriminability statistics (item-total correlations); speededness indicators such as percent item omissions; overt guessing indicators such as longstring item endorsement; and reading grade level.
How do the subtests relate to each other?	3.2	For each subtest, the following statistics were generated: Multivariate correlations and Confirmatory Factor Analysis.
What are the mean score subgroup differences?	3.3	For each subtest, the effect size of mean score subgroup differences were examined using Cohen's d. Female test-takers were compared to male test-takers. Black/African American and Asian test takers were compared to White test-takers. White Hispanic test-takers were compared to White Non-Hispanic test-takers. Finally, Black/African American Hispanic test-takers were compared to Black/African American Non-Hispanic test-takers. Measurement Invariance and variance explained by demographic variables were also explored. Note that throughout the technical report, female test-takers will be referred to as either female test-takers or F; male test-takers will be referred to as either male test-takers or M; Black/African American test-takers will be referred to as either Black/African American test-takers, Black test-takers, or B; White test-takers will be referred to as either White test-takers or W; Asian test-takers will be referred to as either Asian test-takers or A; American Indian/Alaska Native test-takers will be referred as either American Indian/Alaska Native test-takers, AIAN test-takers, or AIAN; Native Hawaiian/Other Pacific Islander test-takers will be referred to as either Native Hawaiian/Other Pacific Islander test-takers, NHPI test-takers, or NHPI; White Hispanic test-takers will be referred as either White Hispanic test-takers or WH; White Non-Hispanic test-takers will be referred to as either White Non-Hispanic test-takers or WnH; Black/African American Hispanic test-takers will be referred as either Black Hispanic test-takers or BH; finally, Black/African American Non-Hispanic test-takers will be referred to as either Black Non-Hispanic test-takers or BnH.
Are the subtests reliable?	3.4	For each subtest, the following statistics were generated: Internal consistency (Cronbach's alpha); and test-retest reliability for T1 (when T1 was administered twice to the same test-takers) and for T2 (when T2 was administered twice to the same test-takers).
Are T1 and T2 equivalent?	3.5	Using a between-subjects design, subtest means for T1 and T2 (first, including all items and then only common items) were compared using independent samples t -tests, under the assumption that assignment to one test or the other was more-or-less random. Accompanying these t -tests were Levene's tests for equality of variances, to determine whether the t -test assumption of homogeneity held. Using a within-subjects design, subtest means for T1 and T2 were examined using paired-samples t -tests. Correlations between the two test forms were also calculated for alternate forms reliability. Linear Mixed Effects Modeling, a technique that partitions variance associated with order effects, form effects, and the interaction between the two, was performed to confirm (or disconfirm) the results of the paired-samples t -test.
Who retests? Do retesters perform better when they retest?	3.6	To discover who is more likely to retest, a series of analyses were performed comparing retesters and non-retesters in terms of their subtest scores and demographics. Additionally, retesters' subset scores on their earlier attempts were compared against the scores on their later attempts.
Was there a test security breach?	3.7	After splitting the first-time test-takers dataset into two groups (earlier group that took the test between January 2015 and August 2016 and later group that took the test between May 2018 and December 2019), means for these two groups were compared using independent samples t -tests.

2.7 Organization of the Report

The results section is organized by the seven aforementioned RQ's that guided AFOQT evaluation at the subtest-level. Each subsection describes the rationale, methodology, and results. The discussion section summarizes the findings and provides recommendation for how to improve the psychometric properties of the cognitive subtests in the future iterations of the AFOQT.

3.0 RESULTS

Note that the results presented throughout this section are for the unstratified sample, unless otherwise stated. The results for the stratified samples are included in the appendices.

3.1 RQ1: What Are the Psychometric Properties of the Subtests?

3.1.1 Descriptives

Descriptive statistics were generated to determine whether the subtest raw scores were normally distributed within and between T1 and T2.

Table 4 summarizes the sample sizes, means, standard deviations, skewness and kurtosis for the raw subtest scores on T1 and T2. The variations in sample sizes within the parallel forms were due to the pairwise deletion of missing values. The variation in sample sizes between the parallel forms was due to the fact that most test-takers were administered T1 rather than T2 and also due to the scrubbing steps that were applied before. Appendix C contains the descriptives stratified by accession source.

Table 4. Unstratified Sample Descriptives

Subtest	T1					T2				
	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
VA	36,098	14.74	4.37	-.19	-.47	32,950	14.58	4.55	-.21	-.52
AR	36,097	13.33	5.00	.11	-.66	32,949	13.17	5.18	.14	-.69
WK	36,091	14.85	5.60	-.08	-1.00	32,945	14.52	5.46	-.06	-.93
MK	36,093	14.18	5.34	.06	-.84	32,943	14.18	5.09	.08	-.75
RC	36,089	17.07	4.03	-.54	-.11	32,938	17.51	4.30	-.69	.13
PS	36,068	10.55	3.97	.12	-.76	32,929	11.03	3.88	.07	-.68
TR	36,064	27.38	6.30	-.19	-.07	32,925	26.84	6.49	-.25	.16
IC	36,058	14.61	6.78	-.29	-.99	32,915	14.57	6.51	-.21	-1.06
BC	36,066	15.23	5.96	.02	-.56	32,926	16.09	5.30	-.03	-.32
AI	36,015	9.42	4.32	.52	-.48	32,878	8.91	3.95	.59	-.17

Note. *N* = Sample Size; *M* = Mean; *SD* = Standard Deviation.

To interpret the meaning of the means and standard deviations within the parallel forms, each subtest's mean should be divided by the total number of points the test-takers could have received for each subtest (which would result in *p*-values). A more detailed discussion about the *p*-values

is covered in greater detail in Section 3.1.2. The comparison of the unstratified raw means and standard deviations between the parallel forms revealed that the subtests in the first half of the AFOQT (verbal and math subtests) were highly similar to one another, but the subtests in the second half of the AFOQT were somewhat different from one another. Specifically, PS and BC seemed easier on T2 than on T1. TR and AI seemed harder on T2 than on T1. The comparison of the stratified samples revealed that for the most part, the accession sources' scores were either slightly higher on T1 or slightly lower (under 1 point). The largest difference was observed for AECP who scored lower on PS on T1 than on T2 ($M = 8.58$ and $M = 9.81$, respectively). USAFA scored lower on BC T1 than on T2 ($M = 13.72$ and $M = 15.00$, respectively). AFRES also scored cadets lower on BC on T1 than on T2 ($M = 15.56$ and $M = 16.61$, respectively). An average difference was 1.19 points, suggesting that some stratification sources found T2 easier than T1.

Recall that negative skewness indicates that the majority of the scores are clustered around the upper end of the distribution and therefore mark an easier test (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 a more difficult test (Tabachnick et al., 2007). The skewness of 0 indicates normal distribution; the skewness ≤ -3 or ≥ 3 is considered extreme (Kline, 2011). The comparison within the parallel forms revealed that RC has the highest negative skew (skewness_{T1} = $-.54$; skewness_{T2} = $-.69$), which indicates that the subtest is least difficult and that there is a potential ceiling effect. AI has the highest positive skew (skewness_{T1} = $.52$; skewness_{T2} = $.59$), which indicates that the subtest is most difficult and that there is a potential floor effect. The floor effect could be due to the fact that the data from those test-takers who are interested in classifying into rated career fields and those who are not were analyzed together. Perhaps, a different result might have emerged if the data were analyzed separately. The comparison of skew between the parallel forms revealed that the subtests seem to be distributed similarly, with VA, WK, RC, TR, and IC showing negative skew or being less difficult, and AR, MK, PS, and AI showing positive skew or being more difficult. BC showed positive skew on T1 ($.02$) and negative skew on T2 ($-.03$), however the difference was small. Section 3.1.2 contains a more detailed discussion of the difficulty parameters between and within the parallel forms.

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 larger sample sizes such as in this study, the impact of negative or positive kurtosis diminishes (Tabachnick et al., 2007). All kurtosis values in this study were within the acceptable range (DeCarlo, 1997).

Overall, the results of the descriptive analyses are consistent with those reported by Carretta et al. (2016) and by Aguilar (2017) suggesting that (1) T1 and T2 are similar in terms of their descriptive statistics (except that some subtests on T2 seemed easier for some stratifications than others) and that (2) some subtests had lower means and positive skew (seemed harder) while others had higher means and negative skew (seemed easier).

3.1.2 Difficulty

CTT statistics were generated for Forms T1 and T2 to further evaluate whether the subtests were of appropriate difficulty.

Table 5 summarizes the minima, maxima, and means for the subtest-level p -values. Recall that in the CTT framework, the p -values serve as an index of difficulty and represent the proportion of test-takers who endorsed the correct responses to the entire subtest, with higher p -values indicating easier tests and lower p -values indicating harder tests (Kline, 2005). The comparison of the p -values within the parallel forms revealed RC ($M_{T1} = .68$; $M_{T2} = .69$) and TR ($M_{T1} = .68$; $M_{T2} = .67$) as the least difficult subtests and AI ($M_{T1} = .47$; $M_{T2} = .44$) as the most difficult subtest. It is possible that AI is difficult for those test-takers who do not have prior exposure to aviation. While a detailed discussion of the items in each subtest is the subject of the item-level technical report (Walsh et al., in press), it is worth mentioning that Carretta et al. (2016) posited an alternate reason for low BC subtest p -values, stating, “There are several items where the blocks touch on their back sides out of the participant’s view. However, none of the example items illustrated this condition” (p. 6). The p -values between T1 and T2 were similar. Appendix D contains the difficulty statistics stratified by accession source.

Table 5. Unstratified Sample Difficulty – p -values

Subtest	T1			T2		
	Min	Max	M	Min	Max	M
VA	.31	.80	.58	.32	.83	.58
AR	.26	.74	.53	.27	.72	.52
WK	.30	.83	.59	.27	.83	.57
MK	.22	.77	.56	.21	.80	.56
RC	.36	.93	.68	.36	.91	.69
PS	.29	.84	.52	.29	.83	.54
TR	.16	.97	.68	.15	.96	.67
IC	.32	.75	.58	.33	.77	.58
BC	.08	.86	.51	.17	.91	.53
AI	.29	.82	.47	.24	.75	.44

Note. Min = Minimum; Max = Maximum; M = Mean.
Total $N_{T1} = 36,099$. Total $N_{T2} = 32,952$. For subtest-specific N , please refer to the Attachment 2.

Overall, the results of the difficulty analyses are consistent with those reported earlier (Aguilar, 2017; Carretta et al., 2016) and suggest that (1) for the most part, T1 and T2 are similar in their difficulty parameters, and that (2) some subtests seemed easier with a large number of easy items while others seemed harder with a lower number of easy items.

3.1.3 Percentiles

Another way to look at difficulty is by examining score percentile distributions. Table 6 summarizes the percentile distribution for Forms T1 and T2. The comparison of the percentile ranks within the forms once again points to RC and TR being the easiest subtests and AI being the hardest subtest. The comparison of the percentile ranks between Forms T1 and T2 showed that the score distributions were very similar for most of the subtests across forms. For eight out of 10 subtests, the percentile score distribution of T1 never departed from that of T2 by more than one

point at any particular percentile. Differences of two points were observed for TR (for the first percentile) and BC (for the fifth, tenth, and twentieth percentiles). Appendix E contains percentiles stratified by accession source. The results for stratified samples were consistent with the unstratified sample. Overall, these results further confirm the conclusion that some subtests are easier than others and are consistent with findings reported by Morath et al. (2008).

Table 6. Unstratified Sample Percentiles

Subtest	T1												
	1	5	10	20	30	40	50	60	70	80	90	95	99
VA	5	7	9	11	12	14	15	16	17	19	20	22	23
AR	3	5	7	9	10	12	13	15	16	18	20	22	24
WK	4	6	7	9	11	13	15	17	19	21	22	24	25
MK	4	6	7	9	11	12	14	16	17	19	22	23	25
RC	6	10	11	14	15	16	18	19	20	21	22	23	24
PS	3	4	5	7	8	9	10	12	13	14	16	17	19
TR	12	17	19	22	24	26	27	29	31	33	36	38	40
IC	1	3	5	8	11	13	15	17	19	21	23	24	25
BC	3	5	7	10	12	14	15	17	19	21	23	25	28
AI	2	3	4	6	7	8	9	10	11	13	16	18	20
Subtest	T2												
	1	5	10	20	30	40	50	60	70	80	90	95	99
VA	4	7	8	11	12	14	15	16	17	19	20	22	23
AR	3	5	6	8	10	12	13	14	16	18	20	22	24
WK	4	6	7	9	11	13	15	16	18	20	22	23	25
MK	4	6	8	9	11	13	14	16	17	19	21	23	25
RC	6	9	11	14	16	17	18	19	20	21	23	23	25
PS	3	5	6	7	9	10	11	12	13	15	16	18	19
TR	10	16	19	22	24	25	27	29	30	32	35	38	40
IC	2	4	5	8	11	13	15	17	19	21	23	24	25
BC	4	7	9	12	13	15	16	17	19	21	23	25	28
AI	2	3	4	5	6	7	8	9	11	12	15	17	19

Note. Total N_{T1} = 36,099. Total N_{T2} = 32,952. For subtest-specific N , please refer to the Attachment 2.

3.1.4 Speededness and Overt Guessing

Table 7 summarizes the percentage of at least one item missing in each set of five items (% Item Omissions) and the percentage of all five items missing at the end of each subtest (% All Last 5 Items Omissions). These statistics serve as indicators of speededness. *Speededness* is defined as “a test characteristic, dictated by a test’s time limit, that results in a person’s test score being dependent on the rate at which items are completed as well as the correctness of the responses” (Carretta et al., 2016, p. 6). Appendix F contains speededness statistics stratified by accession source.

Table 7. Unstratified Sample Speededness

Subtest	T1 (% Item Omissions)								<i>M</i>	% All Last 5 Items Omitted
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40		
VA	.37	.45	.64	2.12	6.13				1.94	.00
AR	.49	1.01	1.25	2.54	5.70				2.20	2.96
WK	.41	.47	.59	1.27	2.55				1.06	1.05
MK	.59	.45	1.02	1.33	3.00				1.28	1.23
RC	.10	.10	.08	.15	.46				.18	.13
PS	.09	.17	.13	.33					.18	.06
TR	.12	.17	.35	1.58	5.91	14.82	25.35	33.42	10.22	29.91
IC	.22	.38	2.27	8.05	16.55				5.49	12.00
BC	.22	.14	.95	4.23	12.18	21.65			6.56	17.55
AI	.15	.16	.25	.95					.38	.33
Subtest	T2 (% Item Omissions)								<i>M</i>	% All Last 5 Items Omitted
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40		
VA	.32	.62	.79	1.80	5.45				1.79	2.46
AR	.68	.86	1.22	2.86	5.83				2.29	3.24
WK	.66	.49	.73	1.24	2.70				1.16	.95
MK	.73	.81	.84	1.82	3.84				1.61	1.83
RC	.11	.11	.10	.20	.42				.19	.12
PS	.08	.14	.12	.30					.16	.05
TR	.12	.17	.43	1.82	6.36	15.47	25.94	33.40	10.46	30.04
IC	.20	.40	2.33	8.26	17.14				5.67	12.54
BC	.17	.11	.44	3.27	11.82	21.90			6.28	17.93
AI	.21	.23	.41	1.16					.50	.39

Note. *M* = Mean. Total N_{T1} = 36,099. Total N_{T2} = 32,952. For subtest-specific *N*, please refer to the Attachment 2.

The comparison within the parallel forms revealed that the prevalence of item omission increased throughout each subtest for the unstratified sample. The same trend was observed throughout the AFOQT itself, with the first half of the test showing smaller omission rates (T1: average 1.33%; T2: average 1.41%) than the second half (T1: average 4.57%; average T2: 4.62%). The subtests with the lowest omission rates were RC (M_{T1} = .18%; M_{T2} = .19%), PS (M_{T1} : .18%; M_{T2} : .16%), and AI (M_{T1} = .38%; M_{T2} = .50%). The subtests with the highest omission rates were TR (M_{T1} = 10.22%; M_{T2} = 10.46%), IC (M_{T1} = 5.49%; M_{T2} = 5.67%), and BC (M_{T1} = 6.56; M_{T2} = 6.28), which is not surprising since these subtests were designed to be speeded. TR contained 40 items and had an allotted time of 7 minutes, which works out to 10.5 seconds per item; IC contained 25 items with an allotted time of 5 minutes (12 seconds per item); and BC contained 30 items with an allotted time of 4.5 minutes (9 seconds per item). The comparison of the omission rates between T1 and T2 revealed a similar pattern. The comparison for the stratified samples revealed that the percent item omissions fluctuates between and within forms across the stratifications, with the largest increase in omissions observed for ANG on TR T2 compared to T1 by 3.77%. See Appendix F.

Another way to evaluate speededness is to look at the proportion of test-takers who omitted all of the last five items on each subtest (both within and between the parallel forms). Consistent with the aforementioned findings for the within-form comparisons, the lowest omission rates were found for RC (M_{T1} = .13%; M_{T2} = .12%), PS (M_{T1} = .06%; M_{T2} = .05%) and AI (M_{T1} = .33%; M_{T2} = .39%). The highest omission rates were found for TR (M_{T1} = 29.91%; M_{T2} = 30.04%), IC (M_{T1}

= 12.00%; $M_{T2} = 12.54\%$), and BC ($M_{T1} = 17.55\%$; $M_{T2} = 17.93\%$). Omission rates between T1 and T2 were similar.

To evaluate whether or not *overt guessing* occurred, we examined the proportion of test-takers who endorsed the same response option (longstring) for the last five items (e.g., A, A, A, A, A) in each subtest. Table 8 summarizes the results for Forms T1 and T2. Comparing within the parallel forms revealed that PS had the lowest rate for its last five items ($M_{T1} = .17\%$; $M_{T2} = .10\%$) and BC had the highest rate ($M_{T1} = 3.15\%$; $M_{T2} = 3.68\%$). When compared to one another, overt guessing rates between Forms T1 and T2 were similar.

Table 8. Unstratified Sample Overt Guessing

Subtest	T1 % Longstring in the Last 5 Items	T2 % Longstring in the Last 5 Items
VA	1.89	1.56
AR	2.51	2.97
WK	.91	1.20
MK	2.16	2.47
RC	.87	.18
PS	.17	.10
TR	2.47	2.82
IC	2.81	2.26
BC	3.15	3.68
AI	.75	.58

Note. Total $N_{T1} = 36,099$. Total $N_{T2} = 32,952$. For subtest-specific N , please refer to the Attachment 2.

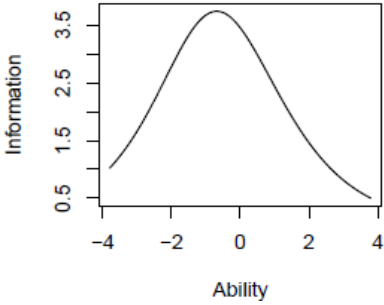
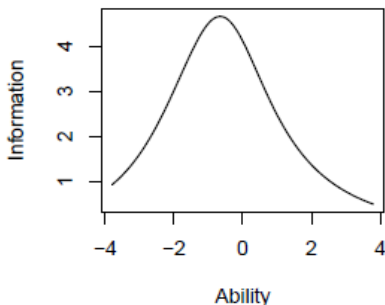
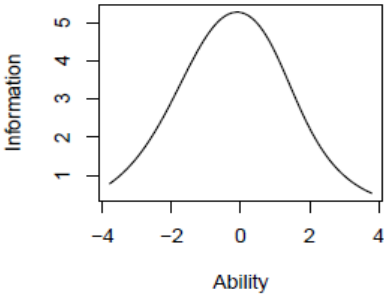
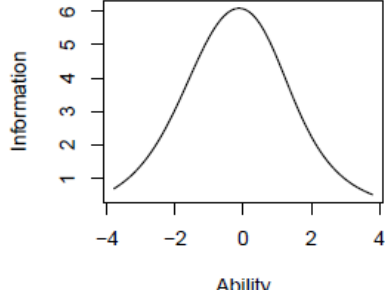
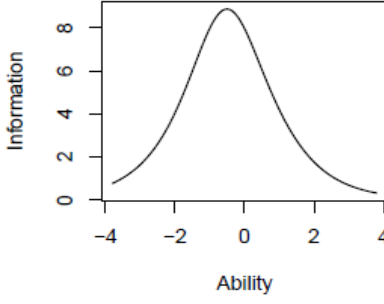
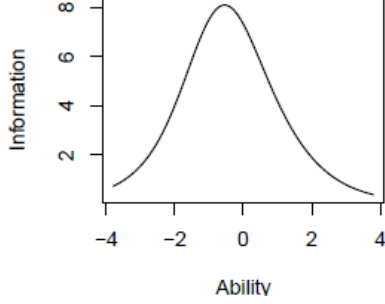
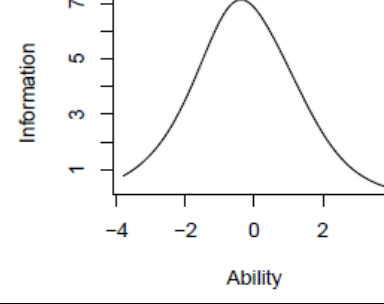
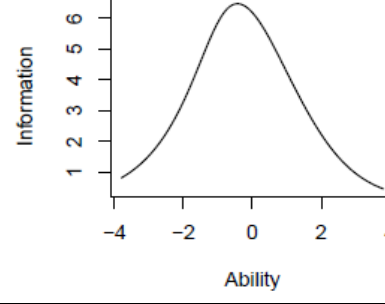
Overall, the results of the speededness and overt guessing are consistent with those reported by Carretta et al. (2016) and suggest that (1) T1 and T2 are similar in terms of their omission rates across all items and across the last five items on the subtests and that (2) omission increases for later items within the subtests and for later subtests within the parallel forms. These results suggest that test-takers either skip or guess on later items.

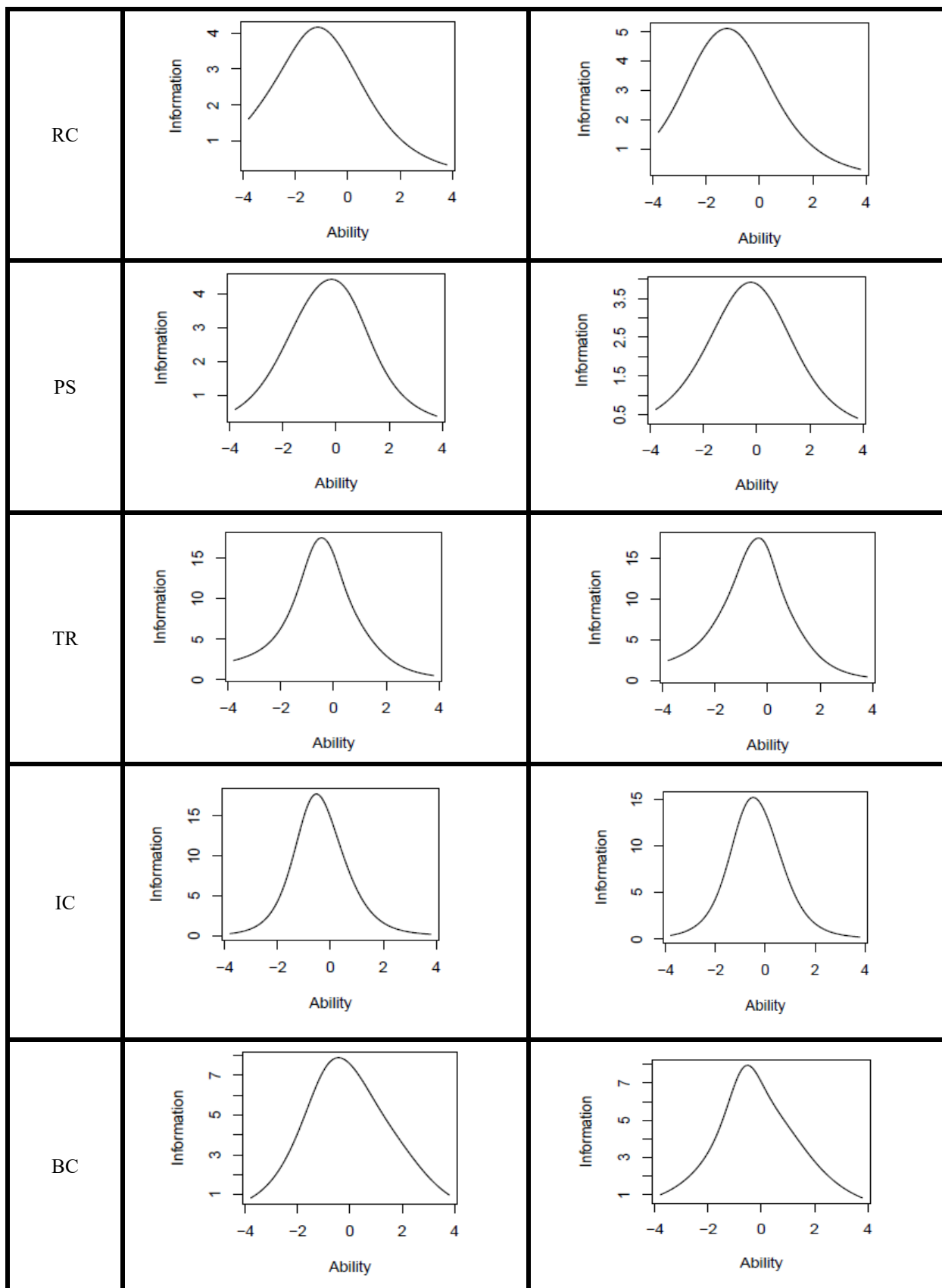
3.1.5 Test Information Function

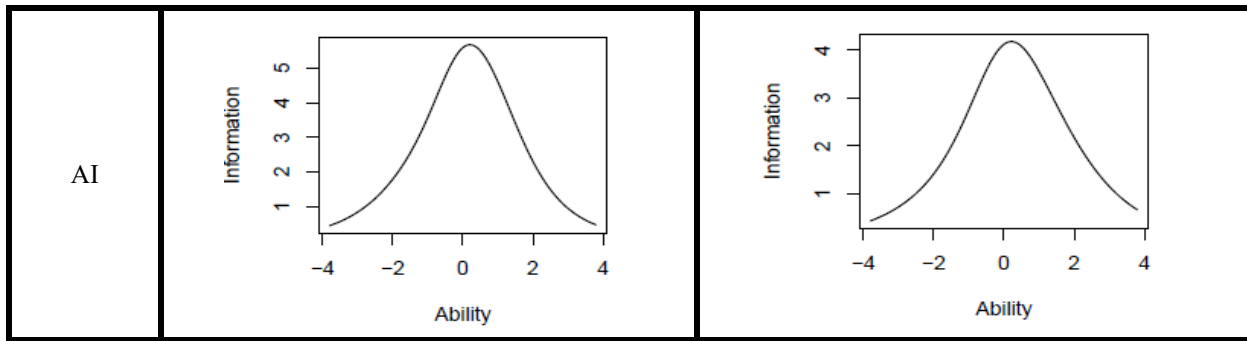
Table 9 displays the Test Information Function (TIF) for each T1 and T2 subtest. Recall that in the IRT framework, *TIF* indicates the relative ability of a scale to discriminate among contiguous trait scores at various locations along the trait continuum. TIF is inversely related to standard error of measurement – which in and of itself represents scale precision (a replacement to the CTT concept of reliability; Ellis & Mead, 2002). In other words, the peak of the distribution marks the range of trait levels that the scale is particularly good at measuring (more precise; Reise & Waller, 2002). For example, if the goal of the selection and classification strategy is to ‘select out,’ poor performers, the measurement precision should be the highest around the lower end of theta. Conversely, if the goal to ‘select in,’ high performers, the measurement precision should be the highest around the higher end of theta. The purpose of the AFOQT is to filter out candidates who do not meet or exceed the minimum cutoff score. Therefore, the TIF distributions should cover lower thetas (approximately between -2.00 and 0). Recall that theta is distributed as a z-score with

a mean of 0 and standard deviation of 1. As can be seen, the peak of most distributions is at or below 0, which indicates that the subtests are fairly easy, with RC being the easiest subtest, but this may be acceptable given the ‘select out’ selection strategy. The AI subtest peaks above 0 which indicates that this is the most difficult subtest. Appendix G contains TIF stratified by accession source.

Table 9. Unstratified TIF

Subtest	T1	T2
VA		
AR		
WK		
MK		





Overall, the results of the TIF analyses for the unstratified sample affirm the common finding that some subtests are easier than others. This information is not available in previous reports. Some differences were observed in TIF graphs across stratified samples, with USAFA's TIF peaking to the left of the other groups especially for AR, MK, and PS, indicating that these subtests appeared easier for USAFA officer candidates relative to other accession sources.

3.1.6 Reading Grade Level

Reading Grade Level (RGL) assesses how easy or difficult a selection of text is to comprehend. The Flesch-Kincaid formula is commonly used to calculate RGL. This method assesses average word and sentence length. Longer words and longer sentences indicate more difficult words, and more difficult syntax, which in turn denote a higher difficulty and RGL (see Figure 1). For example, a RGL of 11 indicates that a passage is best comprehended by individuals with at least 11 grades of basic education. Generally, an eighth-grade RGL (pre-high-school) is the target for employment assessments to ensure that items are understood by approximately 80% of American adults (Kutner et al., 2007).

$$206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right)$$

Figure 1. Flesch-Kincaid Formula

RC RGL Results

Because RC is the most verbally-loaded subtest, it is possible that it would disadvantage the test-takers with lower education compared to those with higher education. As Table 2 shows, nearly one-third of the testing population had completed 13 years of education (T1: 29%; T2: 30%) and another one-third had completed 16 years of education (T1: 30%; T2: 29%). Also, recall that different accession sources take the AFOQT at different times during their careers. For example, OTS-CIV and OTS-AD take the test after they graduate college whereas ROTC take their test during their sophomore year in college (Fedrigo, 2021).

First, we calculated the RC RGL at the reading passage level. The T1 fifth reading passage had the highest RGL at 19.40. The T1 fourth passage had the lowest RGL at 10.00. The average T1

RGL was 14.92. The T2 third passage had the highest RGL at 14.20. The T2 first passage had the lowest RGL at 12.20. The average RGL for T2 was 13.08. Therefore, T1 seems to have a higher variability than T2 but the two forms seem to target at least a high school level of reading.

Next, we correlated the passage-level p -values with the passage-level RGL for RC. To calculate the passage-level p -values, we averaged the p -values for all items in each passage. Although non-significant (with an N of only five RC passages), we found a strong negative correlation between the passage-level RGL and p -values ($r_{T1} = -.74, p > .05$; $r_{T2} = -.81, p > .05$), indicating that as RGL increases, p -values decrease. Therefore, RC passages that require a higher reading level are more difficult (see Table 10).

Finally, we correlated passage-level Cohen's d with the passage-level RGL for RC. As discussed in further detail in Section 3.3.1, the largest RC subgroup differences were found for Black test-takers compared to White test-takers, so we focused on these two subgroups for this exploratory analysis. To generate a passage-level Cohen's d , we averaged the item-level Cohen's d for each passage. With a small N of five RC passages, the results were non-significant. However, the correlations were moderate-to-strong and positive ($r_{T1} = .87, p > .05$; $r_{T2} = .49, p > .05$). Therefore, it seems that as RGL increases, the mean score differences between Black and White test-takers also increase (see Table 10).

Table 10. Unstratified Sample RC RGL Results

Passage	T1				
	RGL	Avg. p -value	r between RGL & Avg. p -value	Avg. B/W Cohen's d	r between RGL & Avg. Cohen's d
1	18.50	.70	-.74	.35	.87
2	11.10	.73		.23	
3	15.60	.60		.28	
4	10.00	.78		.26	
5	19.40	.61		.46	
Passage	T2				
	RGL	Avg. p -value	r between RGL & Avg. p -value	Avg. B/W Cohen's d	r between RGL & Avg. Cohen's d
1	12.20	.78	-.81	.25	.49
2	12.40	.76		.34	
3	14.20	.66		.37	
4	13.50	.60		.43	
5	13.10	.67		.51	

None of the correlations are statistically significant.

Note. RGL = Reading Grade Level; Avg. p -value = Average Passage-Level Difficulty Statistic;
 r = Correlation; Avg. Cohen's d = Average Effect Size.

Overall, the RC RGL results suggest that (1) T1 has more variability than T2, but both forms seem to target at least a high school level of reading; (2) the passages that require a higher RGL are more difficult; and (3) as RGL increases, the mean score differences between Black and White test-takers also increase. With that said, these results are to be interpreted with caution. Future studies should replicate these findings against a larger number of passages.

3.1.7 Discriminability

CTT statistics were generated for Forms T1 and T2 to evaluate whether the subtests can differentiate between low and high ability test-takers.

Table 11 summarizes the minima, maxima, and means for the item-total correlations (ITC's). Recall that in the CTT framework, the ITC serves as an index of discriminability and represents the degree to which the items hang together (measure the same construct), with higher ITC indicating better items and lower ITC indicating poorer items (Kline, 2005). While there is no clear guidance in the literature as to what ITC cutoff to use with some sources recommending a cutoff of .50 (e.g., Kline, 2005), in this study we intentionally used a more liberal cutoff of .30 recommended by Aguinis et al. (2001), such that $ITC \leq .30$ were considered poor and those $> .30$ were considered good. The comparison of the ITC within the parallel forms revealed VA ($M_{T1} = .37$; $M_{T2} = .38$) and RC ($M_{T1} = .37$; $M_{T2} = .40$) as the least discriminable subtests and IC ($M_{T1} = .58$; $M_{T2} = .55$) as the most discriminable subtest. Note that T2 BC had the same mean ITC as T2 RC ($M_{T2} = .40$), indicating that it may be similar to RC in terms of discriminability. The ITC between Forms T1 and T2 were similar. Appendix H contains the discriminability statistics stratified by accession source.

Table 11. Unstratified Sample Discriminability - ITC

Subtest	T1			T2		
	Min	Max	<i>M</i>	Min	Max	<i>M</i>
VA	.24	.48	.37	.28	.54	.38
AR	.28	.54	.42	.27	.57	.43
WK	.33	.61	.47	.27	.60	.46
MK	.31	.56	.45	.29	.59	.43
RC	.23	.49	.37	.25	.53	.40
PS	.22	.59	.42	.25	.54	.41
TR	.17	.68	.41	.23	.67	.42
IC	.32	.67	.58	.32	.64	.55
BC	.31	.55	.44	.21	.60	.40
AI	.31	.61	.45	.29	.57	.42

Note. Min = Minimum; Max = Maximum; *M* = Mean; ITC = Item-Total Correlation; Total $N_{T1} = 36,099$. Total $N_{T2} = 32,952$. For subtest-specific *N*, please refer to the Attachment 2.

Overall, the results of the discriminability analyses are consistent with those reported by Carretta et al. (2016) and by Aguilar (2017). They suggest that (1) Forms T1 and T2 are similar in terms of their discriminability parameters; and (2) some subtests are less discriminable than others and contain numerous problematic items.

3.2 RQ2: How Do the Subtests Relate to Each Other?

3.2.1 Correlations

A *correlation coefficient* (r) is an index of linear association between variables. Correlation coefficients range from -1 to 1, with the sign of the correlation indicating directionality, and the magnitude of the correlation indicating strength of the relationship (Furr & Bacharach, 2008). To examine how each of the cognitive subtests relate to each other, we ran a series of bivariate correlation analyses. We expected to see moderate to strong positive correlations among the subtests due to positive manifold, a phenomenon commonly observed in cognitive ability measurement where the scores on different cognitive tasks show positive inter-correlations (Bartholomew, 2004). We also expected to see strong correlations between more similar subtests (e.g., WK and VA) and weaker correlations between more dissimilar subtests (e.g., TR and VA).

Table 12 displays the zero-order bivariate correlations among the cognitive subtests. The comparison of the correlations within and between the parallel forms for the unstratified sample revealed that all correlations were significant and positive. High correlations were observed among the subtests in the first half of the AFOQT (verbal and math subtests). Specifically, the verbal ability subtests correlated at expected high levels, with VA and WK having the strongest relationship (T1: $r = .68$; T2: $r = .70$), followed by WK and RC (T1: $r = .64$; T2: $r = .64$), and then by VA and RC (T1: $r = .60$; T2: $r = .64$). The quantitative ability subtests (MK and AR) also correlated at expected high levels. Weaker yet significant correlations were observed in the second half of the AFOQT (classification subtests). Specifically, AI and BC were most strongly related to IC. TR was most strongly related to BC. Finally, PS was most strongly related to MK. Further, PS seemed to be more strongly related to the academic subtests than the specialized subtests.

Table 12. Unstratified Sample Bivariate Correlations

Subtest	T1**									
	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI
VA										
AR	.51									
WK	.68	.43								
MK	.45	.70	.36							
RC	.60	.46	.64	.39						
PS	.50	.54	.47	.62	.46					
TR	.30	.42	.20	.34	.25	.21				
IC	.39	.43	.31	.39	.39	.47	.35			
BC	.38	.44	.28	.36	.31	.32	.48	.51		
AI	.35	.33	.34	.28	.39	.43	.25	.57	.35	
Subtest	T2**									
	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI
VA										
AR	.52									
WK	.70	.41								
MK	.47	.74	.36							
RC	.64	.49	.64	.44						
PS	.50	.55	.47	.61	.47					
TR	.29	.40	.20	.35	.28	.21				
IC	.39	.43	.34	.40	.39	.46	.34			
BC	.37	.45	.30	.38	.33	.32	.46	.50		
AI	.33	.31	.33	.28	.35	.43	.24	.58	.33	

**All correlations are significant at $p \leq .010$.

Note. Total $N_{T1} = 36,099$. Total $N_{T2} = 32,952$. For subtest-specific N , please refer to the Attachment 2.

The results of the stratified samples are similar to results presented above. However, the stratified samples seem to have a wider range of correlations. The weakest correlation was observed for the TR-WK relationship on T1 for the USAFA stratification ($r = .09, p \leq .01$). The strongest correlation was observed for the RC-WK relationship on T2 for the AECF stratification ($r = .76, p \leq .01$). Appendix I contains the results of the multivariate correlations stratified by accession source.

Overall, the results of bivariate correlational analyses are consistent with those reported by Carretta et al. (2016). They suggest that (1) Forms T1 and T2 are similar in their correlational magnitudes among the subtest and that (2) the correlations are stronger among the AFOQT verbal and math subtests compared to the specialized subtests.

3.2.2 Confirmatory Factor Analysis

Several CFA models were examined to determine whether or not the factor structure of the current AFOQT Form T cognitive subtests is consistent with the previous findings reported for Form S (Dragow et al., 2010) and for Form T (Carretta et al., 2016). CFA is an analytic technique that relies on *a priori* theory to examine a causal item-to-construct structure (Bryant & Yarnold, 1995). To allow for reasonable comparisons between the current report and the two aforementioned studies, a quick discussion of the methods and results for each is in order.

Drasgow et al. (2010) created item parcels combining a consecutive set of five items together for each subtest to reduce the number of indicators for the AFOQT Form S. The item parcels and subtest scores were used in different models as indicators of higher-order latent constructs. Multiple CFA models were pitted against each other in an attempt to identify the best-fitting model to enhance understanding of the AFOQT data. The models that were compared included single-factor models, four-factor models, five-factor models, and bifactor models. Recall that a bifactor model allows the indicators to load on the general factor and on the specific factors (Drasgow et al., 2010). The results of this investigation led the researchers to conclude that the AFOQT Form S was best described in terms of a bifactor model with five content factors, namely Verbal Ability (VA, WK, GS), Quantitative Ability (AR, MK), Aviation Knowledge (IC, AI, GS), Perceptual Speed (BC, TR), and Spatial Ability (BC, RB, HF)², and with GS and BC allowed to cross-load on multiple content factors. The model had a RMSEA of .05, an NNFI of .98, a CFI of .98, and an SRMR of .06. The researchers also examined orthogonal solutions, which were found to fit the data worse than the oblique solutions.

Caretta et al. (2016) followed a similar method when performing an initial factor analysis of the AFOQT Form T. First, the authors created item parcels for the cognitive subtests. The distinction between this study and the one performed by Drasgow et al. (2010) was in the way item parcels were calculated. Recall that some subtests are speeded (TR, BC) which results in a higher omission rate towards the later items. Therefore, to ameliorate this methodological limitation, Carretta et al. created a randomized set of parcels rather than using first five items, then the next five items, and so on. Next, the item parcels were used as indicators of higher-order constructs. Finally, multiple confirmatory models were pitted against one another. Models with and without cross-loading subtests were run. Orthogonal and oblique solutions were tested as well. The models that were pitted together included a single-factor model, four-factor model, five-factor model, and a hierarchical model with observed variables (parcels) loading onto either four or five content factors, which in turn load onto a single factor. The analyses led the researchers to conclude that a hierarchical factor model fit the data the same way the five-factor model did (RMSEA = .04, NNFI of .90, a CFI of .91, and an SRMR of .05). The five content factors were defined as Verbal Ability (VA, WK, RC, PS), Quantitative Ability (AR, MK), Aviation Knowledge (IC, AI, PS), Perceptual Speed (BC, TR), and Spatial Ability (BC). Note that PS and BC were allowed to cross-load on multiple factors.

The current study sought to combine the best practices from both aforementioned studies. Therefore, as a first step, to statistically identify factor loadings, a randomized set of mutually exclusive and exhaustive parcels were created for all cognitive subtests, resulting in 52 parcels. As a second step, the measurement models for the five content factors were defined as follows: Verbal Ability (VA, WK, RC, PS), Quantitative Ability (AR, MK), Aviation Knowledge (IC, AI, PS), Perceptual Speed (BC, TR), and Spatial Ability (BC). Note that according to this definition, PS and BC were allowed to cross-load. The measurement models for the four content factors were identical to the five factor models except Spatial Ability was removed allowing TR and BC to only load on Perceptual Speed. For a visual depiction of the path diagrams tested in this study see Appendix J. As a final step, we considered the various CFA features. Specifically, because the

² In the AFOQT Form S, there were a few subtests excluded from or modified for the AFOQT Form T. Namely, GS (General Science) became PS; RB (Rotated Blocks) and HF (Hidden Figures) were excluded.

cognitive subtests theoretically measure some facet of cognitive ability, it is likely that they will share some common variance, which would justify the use of an oblique rotation. It is also likely that some items (parcels) may covary within the subtest, therefore permitting the examination of the modification indices. Thus, in our analyses, we tested (1) both the cross-loading models and the non-cross-loading models; (2) orthogonal and oblique solutions; and (3) models with and without covarying the error terms as specified by the modification indices.

We ran a total of 20 CFA models for the cross-loading measurement definition and 20 CFA models for a non-cross-loading measurement definition, using Maximum Likelihood (ML) estimation method³. Consistent with the previous findings, the non-cross-loading models generally performed poorer and therefore the results for these models are not included in this report. Applying the various modification indices to the models did not substantially improve fit and therefore the results for these models are not included in this report either. However, it is worth mentioning that the modification indices recommended covarying parcels within a given subtest (e.g., TR Parcel 3 with TR Parcel 4) rather than between subtests (e.g., TR Parcel 3 with BC Parcel 2). Finally, in line with the previous findings, oblique solutions resulted in an improved fit, albeit minimal. The solution for the Form T2 five-factor model failed to converge due to singularity of the information matrix. Therefore, the results of the oblique analyses are not included in this report as well.

Based on these considerations, Table 13 displays the results of the model-data fit indices for only eight models with cross-loading subtests for Forms T1 and T2. Recall that Drasgow et al. (2010) concluded that the AFOQT Form S cognitive subtests are better described in terms of a bifactor model, whereas Carretta et al. (2016) concluded that the AFOQT Form T cognitive subtests are better described in terms of a hierarchical model or a five-factor model. Based on the results of our analyses, the bifactor model with five content factors performed the best (Form T1: RMSEA = .04, CFI = .93, TLI = .92, SRMR = .04; Form T2: RMSEA = .04, CFI = .93, TLI = .93, SRMR = .04).

³ Similar to the previous literature, the models were estimated using Maximum Likelihood (ML). There are many estimation methods available such as Generalized Least Squares (GLS) and Weighted Least Squares (WLS). However, as research suggests, ML tends to outperform other estimation procedures when the data are non-normal, the sample size is large, and the variables are continuous (Olsson, Foss, Troye & Howell, 2000).

Table 13. Unstratified Sample CFA Model-Data Fit Indices

Model	T1						T2					
	RMSEA	CFI	GFI	AGFI	TLI	SRMR	RMSEA	CFI	GFI	AGFI	TLI	SRMR
Model 1a: g (subtests)	.18	.76	.81	.69	.69	.09	.18	.76	.80	.68	.69	.09
Model 1b: g (parcels)	.09	.55	.53	.49	.53	.09	.09	.56	.55	.51	.54	.09
Model 2: 4 lower-order factors (parcels)	.06	.80	.78	.77	.79	.19	.06	.82	.81	.80	.81	.19
Model 3: 5 lower-order factors (parcels)	.05	.86	.86	.84	.85	.19	.05	.86	.87	.86	.86	.18
Model 4: 2 nd order g and 4 lower-order factors (parcels)	.06	.83	.81	.79	.83	.09	.05	.85	.84	.82	.85	.08
Model 5: 2 nd order g and 5 lower-order factors (parcels)	.04	.89	.89	.88	.89	.08	.04	.90	.90	.89	.90	.08
Model 6: bifactor with g and 4 factors (parcels)	.04	.90	.89	.88	.89	.07	.04	.90	.90	.89	.90	.07
Model 7: bifactor with g and 5 factors (parcels)	.04	.93	.92	.91	.92	.04	.04	.93	.93	.92	.93	.04

Note. g = General Intelligence Factor; RMSEA = Root Mean Square Error of Approximation; CFI = Confirmatory Factor Index; GFI = Goodness of Fit Index; AGFI = Adjusted Goodness of Fit Index; TLI = Tucker Lewis Index; SRMR = Standardized Root Mean Square Residual.

With this, two points are worth mentioning. First, some parcels had negative loadings on some factors. Specifically, for Form T1 Model 6 and 7, PS parcels were negatively loaded on Aviation Knowledge; for Form T2 Model 6, BC parcels were negatively loaded on Perceptual Speed; and for Form T2 Model 7, PS was negatively loaded on Aviation Knowledge. Although undesirable, a similar pattern of results was observed by Drasgow et al. (2010) who found that BC negatively loaded on Spatial Ability. The authors concluded that the presence of negative loadings is likely a result of the magnitude of this subtest's relationship with the general factor. Second, the model-data fit indices are sensitive to model complexity. They tend to be lower in models with more manifest variables (Drasgow et al., 2010). In our case, we have 52 manifest variables, which explains the fit indices being lower than recommended by Hu and Bentler (1998; 1999). Finally, some might argue that there is an inherent statistical bias of fit indices to favor the bifactor model in cognitive ability research and by extension the findings of the current study are inconclusive (Morgan et al., 2015). While this may be the case, recent Monte Carlo simulations showed that fit indices tend to identify the bifactor structures for the data sampled from true bifactor structure and higher-order structure but not for the oblique structure (Morgan et al., 2015). To break the tie between the bifactor and higher-order model when evaluating the model-data fit indices, researchers should use theory to guide the model selection. In the case of the AFOQT Form T, the best fitting model was indeed bifactor. This pattern of results suggests that the current data were sampled from the bifactor structure.

Assuming that the AFOQT Form T is best described by a bifactor model, this further suggests that an important general cognitive ability factor underlies performance on all of the subtests and verbal, quantitative, reasoning, spatial, aircrew, and perceptual speed factors underlie groups of subtests. Specifically, VA, AR, MK, RC, PS, AI, and partially IC load onto a general factor. On

the other hand, WK, TR, and BC load on Verbal Ability, Perceptual Speed, and Spatial Ability, respectively. IC partially loads onto Aviation Knowledge. This is consistent with Carroll's Three Stratum Theory of Intelligence, wherein g and the broad abilities have direct and simultaneous influences on cognitive tasks and are at the same level of inference (Canivez, 2016; Carroll, 1996).

Overall, while there are some variations among the current study and the two previous studies in terms of factor definitions (frequently, operationalized in terms of the available subtests) and CFA models, there are commonalities. Each study: (1) factor analyzed a model with a psychometric g ; (2) factor analyzed the data in terms of four and five content factors; and finally (3) found that the AFOQT data is perhaps best described in terms of g that explains some proportion of variance in the observed variables and in terms of five content factors that also explain some proportion of variance in the observed variables. In terms of practical implication of these findings, it is clear that the cognitive portion of the AFOQT is heavily g -saturated. This likely ensures its criterion-related validity but also results in increased mean score subgroup differences.

3.3 RQ3: What Are the Mean Score Subgroup Differences?

3.3.1 Cohen's d

The USAF strives to achieve diversity and inclusion in its leadership and membership. Therefore, effect sizes were calculated to investigate whether substantial mean score differences exist between racial, ethnic, and gender subgroups. Specifically, we calculated Cohen's d to describe the differences in subtest mean scores between the majority (non-Hispanic) racial group (i.e., White) and the two largest (non-Hispanic) minority racial groups (i.e., Black and Asian); between the majority ethnic group (i.e., White Non-Hispanic and Black Non-Hispanic) and the minority ethnic group (i.e., White Hispanic and Black Hispanic); and between the majority gender group (i.e., male) and the minority gender group (i.e., female). Anyone who identified as multiracial (i.e., belonging to at least two racial categories) was not included in these analyses. Note that the AFOQT demographics data contained two other racial categories including American Indian/Alaska Native and Native Hawaiian/Other Pacific Islander. The sample sizes for these groups were small for Form T1 ($N_{AIAN} = 402$ compared to $N_{White} = 24,859$; $N_{NHPI} = 415$ compared to $N_{White} = 25,828$) and for Form T2 ($N_{AIAN} = 354$ compared to $N_{White} = 22,486$ and $N_{NHPI} = 384$ compared to $N_{White} = 23,481$) in proportion to the majority group. While most of the effect sizes for these racial subgroups were moderate to large, these results should be interpreted with caution due to possible instability of the estimates.

Recall that a common guideline for interpreting Cohen's d in the literature is by evaluating the magnitude of an effect size. Conventionally, the magnitudes of .20 would be considered small; of .50 would be considered moderate; and of .80 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 impact⁴ 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 $d = 0.40$ "

⁴ 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).

(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). See Attachment 3 for more information.

Table 14 summarizes the unstratified statistics for each subtest comparing the gender, racial, and ethnic subgroups. The comparison of the gender effect sizes within the parallel forms revealed that male tended to score higher than female across all subtests. The first half of the test (verbal and math subtests) seems to contain small to moderate effect sizes favoring male test-takers. The second half of the test (classification subtests) seems to contain mostly moderate to large effect sizes favoring male test-takers (except for TR for which the effect size is small for T1 and T2). Specifically, the largest Cohens d were observed for IC and AI, which may be problematic if the USAF desires to attract, develop, and retain female pilots. The comparison between the parallel forms revealed small differences in Cohen's d estimates for each subtest, suggesting that there are similar mean score differences between the two forms. Appendix K contains the results of the mean score subgroup difference analyses stratified by accession source.

Table 14. Unstratified Sample Mean Score Subgroup Differences (Cohen's *d*)

Subgroup	VA						PS					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
F	14.09	4.53		14.00	4.72		8.77	3.54		9.28	3.52	
M	14.97	4.29	.20**	14.79	4.46	.17**	11.18	3.92	.63**	11.65	3.81	.63**
B	12.03	4.22		11.86	4.41		7.95	3.36		8.43	3.40	
W	15.49	4.10	.84**	15.35	4.26	.81**	11.16	3.88	.84**	11.61	3.77	.85**
A	14.07	4.56		14.02	4.66		10.68	3.92		11.21	3.83	
W	15.44	4.11	.33**	15.28	4.28	.29**	11.11	3.88	.11**	11.56	3.77	.09**
WH	14.09	4.34		13.86	4.56		10.10	3.78		10.73	3.70	
WnH	15.58	4.07	.36**	15.44	4.23	.37**	11.19	3.88	.28**	11.63	3.78	.24**
BH	12.38	4.36		12.24	4.45		8.41	3.75		9.23	3.44	
BnH	12.22	4.25	.04	12.02	4.50	.05	8.14	3.45	.08†	8.57	3.50	.19**
Subgroup	AR						TR					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
F	11.44	4.66		11.19	4.82		26.76	6.27		26.23	6.44	
M	14.00	4.94	.53**	13.88	5.12	.53**	27.60	6.29	.13**	27.06	6.49	.13**
B	9.84	4.16		9.58	4.31		23.43	5.99		22.77	6.21	
W	14.12	4.81	.90**	13.95	4.99	.89**	28.27	6.04	.80**	27.71	6.22	.79**
A	14.11	5.27		14.18	5.51		27.47	6.08		27.05	6.43	
W	14.04	4.82	.02	13.87	5.00	.06**	28.20	6.06	.12**	27.64	6.24	.09**
WH	12.44	4.80		12.30	4.84		26.45	6.26		26.10	6.21	
WnH	14.19	4.79	.37**	14.02	5.00	.35**	28.40	5.99	.32**	27.81	6.23	.27**
BH	10.04	4.44		10.27	4.80		23.94	6.27		23.33	6.36	
BnH	10.02	4.21	.00	9.80	4.40	.11*	23.70	6.08	.04	23.13	6.31	.03

Subgroup	WK						IC					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
F	14.09	5.57	.19**	13.36	5.48	.29**	9.74	5.71	1.08**	9.98	5.36	1.05**
M	15.13	5.58		14.93	5.40		16.35	6.26		16.19	6.09	
B	11.81	5.11	.72**	11.39	5.02	.77**	8.95	5.83	1.10**	9.14	5.55	1.10**
W	15.69	5.43		15.40	5.25		15.91	6.37		15.82	6.13	
A	13.68	5.61	.36**	13.22	5.47	.40**	13.09	6.58	.43**	13.28	6.23	.40**
W	15.64	5.44		15.34	5.26		15.84	6.40		15.75	6.15	
WH	14.10	5.50	.32**	13.89	5.32	.31**	13.74	6.61	.36**	13.94	6.26	.32**
WnH	15.82	5.40		15.49	5.24		16.06	6.33		15.92	6.13	
BH	12.59	5.39	.10*	12.02	5.02	.08†	10.26	6.36	.14**	10.94	5.96	.25**
BnH	12.09	5.16		11.59	5.14		9.41	6.11		9.51	5.78	
Subgroup	MK						BC					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
F	12.79	5.12	.36**	12.77	4.92	.38**	13.10	5.77	.50**	14.25	5.19	.48**
M	14.68	5.33		14.69	5.06		15.99	5.84		16.74	5.18	
B	10.99	4.48	.75**	10.93	4.21	.80**	10.56	5.16	1.00**	11.91	4.77	1.01**
W	14.83	5.22		14.83	4.95		16.19	5.70		16.94	5.03	
A	16.26	5.49	.29**	16.02	5.31	.26**	15.12	5.71	.18**	16.07	5.01	.16**
W	14.73	5.22		14.72	4.96		16.12	5.72		16.88	5.06	
WH	13.42	5.08	.28**	13.45	4.86	.28**	14.57	5.76	.30**	15.55	4.99	.29**
WnH	14.85	5.23		14.86	4.96		16.30	5.69		17.01	5.04	
BH	11.16	4.90	.01†	11.53	4.49	.10*	11.64	5.61	.14**	13.30	4.99	.23**
BnH	11.11	4.60		11.08	4.29		10.90	5.31		12.16	4.91	
Subgroup	RC						AI					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
F	15.83	4.07	.42**	16.59	4.59	.29**	7.04	3.34	.79**	6.66	2.89	.82**
M	17.51	3.92		17.84	4.14		10.28	4.31		9.70	3.97	
B	14.40	4.09	.90**	14.51	4.55	.99**	6.63	3.32	.84**	6.37	3.03	.84**
W	17.77	3.72		18.36	3.81		10.15	4.28		9.59	3.93	
A	16.14	4.17	.42**	16.07	4.76	.57**	8.05	3.72	.49**	7.53	3.48	.53**
W	17.73	3.74		18.32	3.83		10.12	4.29		9.58	3.93	
WH	16.67	4.00	.31**	17.00	4.30	.38**	8.64	4.06	.38**	8.24	3.65	.37**
WnH	17.84	3.70		18.45	3.77		10.27	4.29		9.69	3.93	
BH	14.85	4.21	.05	15.21	4.78	.11*	6.99	3.44	.04	6.77	3.08	.07
BnH	14.65	4.12		14.69	4.59		6.86	3.50		6.56	3.17	

† $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p . Underlined are Cohen's d for which minority test-takers outperformed majority test-takers.

Note. F = Female; M = Male; B = Black/African American; A = Asian; W = White; WH = White Hispanic; WnH = White Non-Hispanic; BH = Black/African American Hispanic; BnH = Black/African American Non-Hispanic; M = Mean; SD = Standard Deviation; d = Cohen's d . Total T1: $N_{\text{Female}} = 9,497$, $N_{\text{Male}} = 26,565$; Total T2: $N_{\text{Female}} = 8,618$, $N_{\text{Male}} = 24,286$; Total T1: $N_{\text{Black}} = 3,290$, $N_{\text{White}} = 25,419$; Total T2: $N_{\text{Black}} = 3,019$, $N_{\text{White}} = 23,084$; Total T1: $N_{\text{Asian}} = 2,142$, $N_{\text{White}} = 25,003$; Total T2: $N_{\text{Asian}} = 1,935$, $N_{\text{White}} = 22,802$; Total T1: $N_{\text{White Non-Hispanic}} = 24,064$, $N_{\text{White Hispanic}} = 3,440$; Total T2: $N_{\text{White Non-Hispanic}} = 21,875$, $N_{\text{White Hispanic}} = 3,044$; Total T1: $N_{\text{Black Non-Hispanic}} = 4,303$, $N_{\text{Black Hispanic}} = 541$; Total T2: $N_{\text{Black Non-Hispanic}} = 3,916$, $N_{\text{Black Hispanic}} = 500$. For subtest-specific N , please refer to the Attachment 2.

The comparison of the B/W effect sizes within the parallel forms for the unstratified sample revealed that White test-takers tended to score higher than Black test-takers across all subtests. Both halves of the test have fairly large Cohen's *d*, ranging from .72 (WK) to 1.10 (IC) on T1 and from .77 (WK) to 1.10 (IC) on T2. The comparison between the parallel forms revealed small differences in Cohen's *d* estimates for each subtest, suggesting that there are similar mean score differences between the two forms.

The comparison of the A/W effect sizes within the parallel forms revealed that White test-takers tended to score higher than Asian test-takers across eight subtests. The Asian subgroup outscored White test-takers on two subtests, namely AR and MK. Both halves of the AFOQT seem to have small to moderate Cohen's *d*, ranging from -.02 (AR) to .49 (AI) on T1 and from -.06 (AR) to .57 (RC) on T2. The comparison between the parallel forms revealed small differences in Cohen's *d* estimates for each subtest, suggesting that there are similar mean score differences between the two forms.

The comparison of the two ethnic subgroups within the parallel forms revealed that (1) White Non-Hispanic test-takers tended to score higher than White Hispanic test-takers across all subtests and (2) Black Hispanic test-takers tended to score higher than Black Non-Hispanic test-takers across all subtests. All effect sizes were small between and within the parallel forms.

Overall, the results of the aforementioned analyses are consistent with those reported by Morath et al. (2008) and suggest that (1) Forms T1 and T2 are similar in terms of their mean score subgroup differences (with some accession sources showing either an increase or decrease in effect sizes between T1 and T2); and that (2) the cognitive subtests seem to be biased against female-, Black, and Asian test-takers. For a detailed description of the problematic items that may drive these differences, please see Walsh et al. (in press).

Because test-takers vary across accession sources in ways that may affect the interpretation of any subgroup differences, Cohen's *d* calculations were repeated within each accession source. To test whether or not the subgroup differences found across the unstratified sample are a function of accession source differences, we first examined the demographic representation of minorities in each accession source (see Appendix A). Note that the AECOP population is fairly small with some sample sizes falling below 10 test-takers per subgroup (e.g., T1: $N_{\text{Black-Hispanic}} = 5$; T2: $N_{\text{Black-Hispanic}} = 2$), therefore their results will not be discussed in this report.

There were fewer female test-takers in OTS-AD (T1: 21%; T2: 19%) and more in OTS-CIV (T1: 30%; T2: 29%) and in ROTC (T1: 28%; T2: 29%). However, these values are on par with the general test-taking gender distribution, with female test-takers making up roughly 30% (see Table 2). There was large variability in the Black and White test-taking distribution across the accession sources. Specifically, there are fewer Black test-takers in the USAFA (T1: 10%; T2: 9%) and more in the AFRES (T1: 20%; T2: 19%) compared to the general test-taking population of 14% (see Table 2). Findings for Asian-White test-takers are similar to Black-White test-takers in that there are fewer Asian test-takers in the ANG (T1: 7%; T2: 7%) and more in ROTC (T1: 12%; T2: 11%) and in USAFA (T1: 11%; T2: 11%) compared to the general test-taking population of 10% (see Table 2). Finally, the distribution of Hispanic test-takers compared to Non-Hispanic test-takers across the accession sources revealed that the former group is less represented in the USAFA (T1:

11%; T2: 11%) and in the ANG (T1: 11%; T2: 10%). There are more Hispanic test-takers in ROTC (T1: 16%; T2: 17%) and in AFRES (T1: 16%; T2: 15%). The general population has 14% of Hispanic test-takers (see Table 2).

In Morath et al. (2008), Black, Hispanic, and female test-takers were more likely to have been ROTC scholarship applicants, while White and male test-takers were more likely to have been USAFA cadets. The authors note that these qualitative sample differences are meaningful, because the ROTC test-takers attend a variety of education institutions across the nation, each with different admission and curriculum requirements. The USAFA, on the other hand, sets high admission criteria with a fairly standardized curriculum. Therefore, an over-representation of racial, ethnic, and gender minorities among some accession sources and under-representation of minorities in other accession sources could exacerbate subgroup differences when the entire sample is evaluated.

Indeed, in the Morath et al. study, there was a larger difference between the stratified sample and the unstratified samples. For example, male test-takers made up 75% of the Form S1 unstratified sample and 90% of the Form S1 stratified (USAFA) sample. Similarly, Black test-takers made up 11% of the Form S1 unstratified sample and only 3% of the Form S1 stratified (USAFA) sample. The present study leveraged a larger dataset and therefore, the demographics statistics were a bit more balanced. For example, male test-takers made up 74% of the Form T1 unstratified sample and 74% of the T1 stratified (USAFA) sample. Similarly, Black test-takers made up 14% of Form T1 unstratified sample and 10% of the T1 stratified (USAFA) sample. Due to these fairly large qualitative differences between Morath et al. and our study, we will not make stratified comparisons between Form S and Form T for this section.

The comparison of gender mean score difference across the cognitive subtests within the parallel forms revealed that (1) all accession sources had roughly an equal number of subtests with small and moderate effect sizes and at least one subtest with a large effect size; (2) all accession sources revealed large effect sizes for IC and some for AI (T1: OTS-AD, ROTC, ANG, and AFRES; T2: OTS-AD, ROTC, ANG, and AFRES); and (3) male test-takers always outscored their female counterparts. There was a high degree of convergence when comparing the effect sizes between the parallel forms.

Comparison of the Black and White mean score difference across the cognitive subtests within the parallel forms revealed that (1) all accession sources had either moderate or large effect sizes (no small effect sizes); (2) all accession sources had large effect sizes for IC on both Forms T1 and T2 and for RC on T1; and (3) White test-takers always outscored their Black counterparts. There was a high degree of convergence when comparing the effect sizes between the parallel forms.

The comparison of the Asian/White mean score difference across the cognitive subtests within the parallel forms showed that (1) all accession sources had a large proportion of subtests with small effect sizes, some with moderate effect sizes, and none with large effect sizes; (2) most accession sources had moderate effect sizes for AI (except $T1_{OTS-CIV} = .39$, which is arguably approaching the moderate effect size threshold of .40); and (3) White test-takers outscored their Asian counterparts on most of the subtests, except for AR and MK, on which Asian test-takers scored higher than White test-takers across all accession sources except for Form T1 OTS-AD, ANG, and

AFRES. There was a high degree of convergence when comparing the effect sizes between the parallel forms.

Overall, the results of the subgroup analyses across the accession sources pointed to the fact that while USAFA had a lower proportion of minorities than the ROTC, the difference in the present study was smaller than that reported by Morath et al. (2008). With that, USAFA showed slightly lower gender and Black and White subgroup differences than did ROTC. This suggests that subgroup differences exist between accession sources. However, these analyses do not completely explain the subgroup differences observed in the combined sample.

3.3.2 Measurement Invariance

Several CFA and chi square tests were run to evaluate the AFOQT Form T data for MI. Recall that MI analysis shows whether individuals with the same latent trait but from different groups have the same probability of endorsing the correct response (Waller et al., 2000).

There are four MI models: Configural, metric, scalar, and strict. *Configural invariance* is established when the measure has the same configuration (factor structure) for both groups (Vandenberg & Lance, 2000). Conversely, configural variance is established when the two groups have either a different number of indicators or different number of latent traits. After configural invariance is established, metric invariance is tested.

Metric invariance constrains the factor loadings for both groups to be the same (Vandenberg & Lance, 2000). The only things that can vary are the intercepts and the residuals. The limitation of the metric invariance is that it does not compare the groups. Rather, it allows a researcher to state that the construct is measured the same way across both groups. After metric invariance is established, scalar invariance is tested.

Scalar invariance constrains the factor loadings and the intercepts for both groups to be the same (Vandenberg & Lance, 2000). The only things that can vary are the residuals. The scalar invariance model is the most desired in practice because it allows for the comparison of the latent constructs between groups. Note that for model identification, one group's latent mean need to be constrained to 0 whereas the latent mean for the other group would be allowed to vary. For example, if we constrain the means of female test-takers to 0 and the mean of male test-takers to .15, this would indicate that male test-takers have higher academic achievement. Conversely, if male test-takers' mean is -.3, it would indicate that male test-takers have lower academic achievement.

Note that there is a fourth kind of invariance called residual invariance which is considered to be the strictest form of invariance and is rarely achieved in practice. While we did test for strict invariance, we do not report the results here. Also note that these analyses were not run on the stratified samples.

Table 15 summarizes the CFA model-data fit indices (to be interpreted similarly to the standard CFA indices; Hu & Bentler 1998, 1999) for gender and two racial groups. Because the largest mean score differences were observed for these subgroups, the MI analyses were not run on the rest of the racial and ethnic subgroups. Note that our starting model was bifactor and consisted of

a factor representing g and five content factors. Goodness-of-fit statistics were examined at each step to assess the extent to which adding this constraint degrades fit. As can be seen, the changes in fit indices are either small or nonexistent when the constrained models (i.e., metric and scalar) are compared to the baseline (i.e., configural) solution. CFI on Forms T1 and T2 changed trivially (from .92 to .91) and TLI on T2 changed trivially as well (from .92 to .91). In sum, the results suggest that there is little or no differential item and test functioning across minority and majority groups.

Table 15. MI Model-Data Fit Indices

Subgroup	Models	T1				T2			
		CFI	TLI	RMSEA	SRMR	CFI	TLI	RMSEA	SRMR
Female/Male	Configural	.92	.91	.04	.05	.92	.92	.04	.05
	Metric	.91	.91	.04	.05	.92	.92	.04	.05
	Scalar	.91	.91	.04	.05	.91	.91	.04	.05
Black/White	Configural	.91	.91	.04	.05	.92	.91	.04	.05
	Metric	.91	.91	.04	.05	.91	.91	.04	.05
	Scalar	.91	.91	.04	.05	.91	.91	.04	.05
Asian/White	Configural	.91	.91	.04	.05	.92	.91	.04	.05
	Metric	.91	.91	.04	.05	.92	.91	.04	.05
	Scalar	.91	.91	.04	.05	.91	.91	.04	.05

Note. CFI = Confirmatory Factor Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

Although trivial difference in model-data fit indices were observed, we ran additional tests. Within the CFA framework, the researchers want to accept the null hypothesis that the model fits the data. To do so, a Maximum Likelihood fit function is applied to each nested model (recall that the metric model is nested in the configural model, the scalar model is nested in the metric model, etc.) and the fit indices are examined. Some of the fit indices include chi square tests which are notoriously sensitive to large sample sizes and are often considered badness of fit indices since a significant p -value indicates lack of fit.

As Appendix L shows results of the measurement invariance chi-square tests, and as expected, most of the p -values were significant at $p < .001$ level, which is not desirable because it suggests that the more constrained models are worse than the preceding less constrained models. This may suggest that there is no MI in the dataset. However, as was mentioned before, the chi square tests are sensitive to large sample sizes and therefore may result in Type I error. In sum, the statistical significance should be interpreted with caution.

Overall, this pattern of results is consistent with previous MI investigations (e.g., Drasgow et al., 2010) and suggests that the AFOQT Form T cognitive subtests function similarly across the gender and racial subgroups.

3.3.3 Variance Explained by Demographic Variables

As stated in previous sections, demographics may impact cognitive ability test scores. As an additional way of exploring these relationships, we performed stepwise regression analyses using demographic variables as predictors and subtest scores as outcomes. Stepwise regression is empirically driven, and therefore is not recommended in any instance where the model could be reasonably driven via theory. However, as this was purely an exploratory analysis with no underlying theory, we decided to use both forward and backward stepwise regression analyses. Note that these analyses were not performed for the stratified samples.

Stepwise regression adds each predictor into the model one at a time, with the first added variable explaining the most variance in the outcome, and later variables explaining less variance (Thorndike, 1982). We set the “stopping point” of the procedure at $p \leq .01$, so if a predictor did not explain a significant amount of variance at the level specified, the model would stop running and be considered complete. Further, if a variable added later makes a previously added variable non-significant, it can be removed from the model. Table 16 shows the results of the final models.

Table 16. Unstratified Sample Stepwise Regression Results

VA				PS			
Step	Variable	R^2	ΔR^2	Step	Variable	R^2	ΔR^2
1	Black	6.00%			Age		
2	HispW.	6.80%	.80%	1	Black	18.40%	
3	Asian	7.40%	.60%		Gender		
4	NHPI	7.70%	.30%	2	HispW.	18.80%	.40%
5	AIAN	7.90%	.20%	3	NHPI	18.90%	.10%
6	Gender	8.10%	.20%	4	AIAN	19.00%	.10%
				5	Asian	19.10%	.10%
AR				TR			
Step	Variable	R^2	ΔR^2	Step	Variable	R^2	ΔR^2
1	Black	11.20%		1	Black	5.80%	
	Gender			2	Age	6.50%	.70%
2	Age	13.00%	1.80%	3	HispW.	7.10%	.60%
3	HispW.	13.80%	.80%	4	NHPI	7.20%	.10%
4	NHPI	14.00%	.20%	5	AIAN	7.20%	.00%
5	AIAN	14.10%	.10%	6	Asian	7.30%	.10%
6	Asian	14.10%	.00%	7	Gender	7.30%	.00%
WK				IC			
Step	Variable	R^2	ΔR^2	Step	Variable	R^2	ΔR^2
1	Black	4.70%		1	Black	25.80%	
2	Age	6.30%	1.60%		Gender		
3	Asian	7.00%	.70%	2	Asian	26.70%	.90%
4	HispW.	7.70%	.70%	3	Age	27.20%	.50%
5	Gender	8.10%	.40%	4	HispW.	27.70%	.50%
6	AIAN	8.40%	.30%	5	NHPI	27.80%	.10%
7	NHPI	8.60%	.20%	6	AIAN	27.90%	.10%
MK				BC			
Step	Variable	R^2	ΔR^2	Step	Variable	R^2	ΔR^2
1	Age	15.00%		1	Black	11.70%	
	Black				Gender		
2	Gender	17.40%	2.40%	2	Age	12.70%	1.00%
3	Asian	18.00%	.60%	3	HispW.	13.10%	.40%
4	HispW.	18.50%	.50%	4	Asian	13.20%	.10%
5	AIAN	18.60%	.10%	5	NHPI	13.30%	.10%
	NHPI	18.70%	.10%	6	AIAN	13.40%	.10%
RC				AI			
Step	Variable	R^2	ΔR^2	Step	Variable	R^2	ΔR^2
1	Black	7.20%		1	Black	15.60%	
2	Gender	8.80%	1.60%		Gender		
3	Asian	10.00%	1.20%	2	Aian	17.00%	1.40%
4	HispW.	10.70%	.70%	3	HispW.	17.80%	.80%
5	NHPI	10.90%	.20%	4	AIAN	17.90%	.10%
6	AIAN	11.10%	.20%	5	NHPI	18.10%	.20%
				6	Age	18.10%	.00%

Note. R^2 = Percent of Variance Explained; ΔR^2 = Change in R^2 when a variable is added or removed from a regression; Black = Black/African American; HispW. = White Hispanic; AIAN = American Indian/Alaska Native; NHPI = Native Hawaiian/Other Pacific Islander. Total N = 69,051.

The following variables seem to drive the most variance in subtest scores: Black, Age, Asian, and Gender. The subtests with the most variance explained by demographics are IC (27.9%), followed by PS (19.1%), MK (18.7%), and AI (18.1%). The subtests with the least amount of variance driven by demographics include TR (7.3%), VA (8.1%), and WK (8.6%).

3.4 RQ4: Are the Subtests Reliable?

Two methods were used to evaluate the reliability of the subtests: 1) internal consistency; and 2) test-retest reliability (i.e., test-takers took the same form twice). Results of each are presented in the sections below.

3.4.1 Internal Consistency

To measure internal consistency, we calculated Cronbach's alpha for each subtest. *Cronbach's alpha*, or "raw" alpha is the most widely utilized measure of internal consistency (Cronbach, 1951). It uses the covariance among items and the overall variance of the scale to produce a value between 0 and 1, with higher values indicating better internal consistency. The standards for acceptable values of alpha range across studies and fields. Taber (2018) performed a critical review of Cronbach's alpha and indicated, among other conclusions, that .70 is the most widely used threshold for acceptable internal consistency. At the same time, alphas at or above .90 are considered redundant (Streiner, 2003).

Table 17 displays the alphas for each subtest. Comparing the values between and within the parallel forms reveals that all subtests met or exceeded the .70 threshold. For Form T1, the values ranged from .73 (RC) to .92 (IC) and for T2 from .74 (PS) to .91 (IC).

Table 17. Unstratified Sample Internal Consistency (Cronbach's Alpha)

Subtest	T1	T2
VA	.74	.75
AR	.80	.82
WK	.85	.84
MK	.83	.82
RC	.73	.78
PS	.75	.74
TR	.89	.89
IC	.92	.91
BC	.86	.82
AI	.80	.75

Note. Total N_{T1} = 36,099. Total N_{T2} = 32,952.
For subtest-specific N , please refer to the Attachment 2.

Internal consistency was also computed for the stratified samples. Although most of the results for the stratified samples were consistent with the ones reported above, there were a few that did not meet the .70 threshold. Specifically, on Form T1, OTS-AD, AECP, and USAFA had alphas equal

to .66, .66, and .69, respectively for PS. Additionally USAFA had an alpha of .67 for the VA subtest. Form T2 showed similar differences for the stratifications, such that for subtest PS, OTS-AD, AECF, USAFA, and ANG, all had alphas of .63, .67, .66, and .68, respectively. USAFA again had a lower internal consistency for the VA subtest (.69). Appendix M contains the results of the internal consistency analyses stratified by accession source. The results for the stratified samples were consistent with the above results.

Overall, the results of the internal consistency analyses are consistent with those reported by Carretta et al. (2016) and by Aguilar (2017) and suggest that T1 and T2 are similar in their internal consistencies. A few stratifications showed lower than desirable internal consistencies.

3.4.2 Test-Retest Reliability

AFOQT Retesting Policy

The AFOQT retest policy allows for an initial test and two retests with 150 days in-between retests (or 180 days, according to an older policy; Fedrigo, 2021). Note that a third attempt is allowed if a waiver is granted. According to the same policy, the test-takers are supposed to retest on the parallel forms. For example, if they took T1 on the first attempt, they should take T2 on their second attempt, followed by T1 on their third attempt. While this is usually true, there are some instances when the test-takers received the same form twice (e.g., T1 on their first and second attempt). Therefore, the dataset for the test-retest analyses included those test-takers who took T1 twice and those who took T2 twice, sequentially and non-sequentially, and correlating their earlier scores with their later scores. Sometimes the test scores from first attempt were correlated with the third attempt, other times from first and second, and so on.

Test-Retest Reliability Results

Recall, that *test-retest reliability* is the most common CTT statistic for determining whether an assessment measures the same construct across the two testing occasions (Furr & Bacharach, 2008). Any deviation from a perfect correlation (< 1.00) are due to (1) measurement error; (2) true score change; or (3) a combination of both (Furr & Bacharach, 2008; Wainer et al., 2000).

Table 18 summarizes the means, standard deviations, and test-retest reliabilities for each subtest on Forms T1 and T2. The sample size was 714 for T1 and 564 for T2. The sample sizes fluctuated slightly based on the pairwise deletion of missing values. The comparison of the means and standard deviations within the parallel forms for the unstratified sample indicates that the test-takers scored higher across all subtests on their later attempt, ranging from .18 points to 1.91 points on T1 and from .18 points to 1.73 points on T2. As the literature suggests, the time lag between the retests should be taken into account because if it is too short, any improvements in the test scores may be due to practice effects (Thorndike, 1982). However, in the case of the AFOQT with the time lag of at least 150 days and the fact that the test-takers have no knowledge as to what form they are going to be administered when they retest, the practice effects are unlikely. A more likely explanation for the increase in the scores is a combination of factors including measurement error and true score change (due to extra studying). The comparison of the means and standard deviations between the parallel forms revealed a similar pattern of results.

Appendix N contains test-retest statistics stratified by accession source. Most test-retest reliabilities were acceptable, with a few exceptions. First, there is a high variability in test-retest reliabilities for USAFA on BC for T2 than on T1 (the lowest $r_{T2T2} = .47$, p ns). This is likely due to the fact that USAFA test-takers are least motivated to retest which is evident in the sample sizes for T1 of 82 people and for T2 of 8 people. Overall, the sample sizes for the accession sources are generally low which may explain the fluctuations in the test-retest reliabilities. Of all subtests, TR emerged as having the lowest test-retest reliability across the accession sources.

Table 18. Unstratified Sample Test-Retest Reliability

Subtest	T1-T1 (same form)					T2-T2 (same form)				
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	r_{T1-T1}^{**}	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	r_{T2-T2}^{**}
VA	12.24	4.24	13.65	3.89	.79	11.65	4.59	13.38	4.13	.76
AR	10.95	4.52	11.77	4.48	.82	9.91	4.45	10.76	4.39	.80
WK	11.34	5.26	12.38	5.23	.86	11.31	5.41	12.21	5.23	.87
MK	11.86	4.86	12.95	4.60	.82	10.97	4.29	12.05	3.96	.76
RC	14.85	4.26	15.92	4.05	.78	14.66	4.70	15.78	4.44	.79
PS	9.09	3.51	9.28	3.54	.84	9.01	3.32	9.19	3.36	.83
TR	26.17	6.40	27.63	6.60	.73	24.49	6.65	25.99	7.20	.67
IC	12.63	6.85	14.53	6.96	.82	12.14	6.30	13.41	6.54	.84
BC	13.72	5.77	15.14	5.96	.80	14.21	5.07	15.70	5.12	.81
AI	8.47	4.41	9.26	4.59	.88	7.64	3.77	8.12	4.07	.85

**All correlations are significant at $p \leq .010$.

Note. M_{Earlier} = Mean of Earlier Test Administration; SD_{Earlier} = Standard Deviation of Earlier Test Administration; M_{Later} = Mean of Later Test Administration; SD_{Later} = Standard Deviation of Later Test Administration; r = Correlation. Total $N_{T1-T1} = 714$. Total $N_{T2-T2} = 564$. For subtest-specific N , please refer to the Attachment 2.

The comparison of test-retest reliabilities within the parallel forms for the unstratified sample showed that for the most part they were at acceptable levels ($r \geq .70$; Nunnally, 1978). The strongest reliability was found for AI on T1 ($r = .88$, $p < .01$) and for WK on T2 ($r = .76$, $p < .01$). The weakest reliability for both parallel forms was found for TR, with T1 TR reliability being .73 and T2 TR reliability being .67. Typically, reliabilities below .70 are considered evidence of poor measurement and may point to the problematic areas (items, subtests). The mean test-retest reliability coefficients for the 10 subtests on T1 and T2 were found to be .81 and .80, respectively. While the comparisons between the parallel forms revealed a similar pattern of results, MK (.06 points) and TR (.07 points) showed the greatest amount of difference in their test-retest reliabilities. The comparison for the stratified samples showed that test-retest reliability fluctuated somewhat between T1 and T2 across the accession sources, with the largest difference observed for USAFA for whom the T2-T2 correlations were much lower than the T1-T1 correlations by an average of .26 units.

Note that since the previous Form T technical reports do not report test-retest reliabilities, we compared our findings against those reported by Morath et al. (2008) for the AFOQT Form S. As mentioned before, due to some qualitative changes between Form S and Form T as well as due to rather modest sample sizes used in calculating of the Form S test-retest reliabilities (S1: $N = 67$; S2: $N = 50$), these comparisons are to be interpreted with caution. For the most part, the pattern of

results was similar between the current study and the one performed by Morath et al., with: (1) TR showing the lowest reliability (Form S1: $r = .64$, p not reported; Form S2: $r = .63$, p not reported); (2) Form T reliabilities being higher than those for Form S; and (3) the differences between the reliabilities being smaller for Form T (up to 7 points on TR) than for Form S (up to 12 points on VA), suggesting measurement improvement for Form T.

3.5 RQ5: Are T1 and T2 Equivalent?

In addition to internal consistency and test-retest reliability, *alternate forms reliability* was computed. It serves as a way to determine how scores on one form correlate with the scores on the other form (Furr & Bacharach, 2008). It can also serve as a way to determine the equivalency of the parallel test forms (Furr & Bacharach, 2008). Appendix O contains the results of the stratified samples between-subjects alternate forms reliability analyses across all items. Appendix P contains the results of the stratified samples between-subjects alternate forms reliability analyses across common items only. Finally, Appendix Q contains the results of the stratified samples within-subjects alternate forms reliability analyses across all items.

3.5.1 Methodology

Alternate forms reliability was assessed in four ways. First, between-subjects mean differences across all items were examined between Forms T1 and T2 using independent samples t -tests. For this analysis, the order of assessment (i.e., T1 then T2 vs. T2 then T1) was balanced so the sample sizes were equal. With this method, differences could be attributed to non-random sampling, confounding the sample and the difference in forms. Although the AFOQT is believed to be randomly sampled, this should still be kept in mind when evaluating results.

Second, between-subjects mean differences across common items only were examined between Forms T1 and T2 using independent samples t -tests. This helps to eliminate the form-sample confound which could be present in the first method. However, this method introduces a new source of error – item location. While a detailed discussion of the common items is the subject of the item-level analyses (Walsh et al., in press), it is worth mentioning that for the purposes of the present analysis, common items were defined as the ones having identical stems and response options between Forms T1 and T2. As previously discussed, common items did not appear in the same order on both assessments, therefore, results should be interpreted with caution due to the potential order effects.

Third, within-subjects mean differences across all items were examined for the test-takers who took Form T1 and then retested on T2 or who took T2 and then retested on T1 using paired samples t -tests. For this analysis, the order of assessment (i.e., T1 then T2 vs. T2 then T1) was balanced so the sample sizes were equal. This step controls for the potential practice effects that come with taking a test more than once. For analyses with unequal variance as indicated by Levene's test, Welch's t -test was used instead of the traditional Student's t -test (Levene, 1961).

Our fourth method was a more statistically advanced way of exploring alternate forms reliability: Linear Mixed Effects Models (LMEM). *LMEM* are mathematical models with fixed and random effects. This could include hierarchical regression, hierarchical linear modeling, multilevel regression, multilevel linear modeling, linear mixed models, and random coefficient models. These

models allow for one or more sources of non-independence of effects. Therefore, LMEM can partition variance associated with order effects, test form effects, and the interaction between the two (Brauer & Curtin, 2018). Note that this analysis was not run for the stratified samples.

3.5.2 Between-Subjects Alternate Forms Reliability Results

Table 19 displays the results of the between-subjects alternate forms reliability analysis across all items. Table 20 displays the results of the same analyses but across common items only (note that RC did not have common items between T1 and T2). The results between the two types of analyses were consistent for the stratified and unstratified samples. Specifically, the tests using all items showed two subtests with non-significant mean differences (MK and IC). The tests using only the common items showed one subtest (WK) with non-significant differences and three subtests (AR, MK, and BC) with differences significant only at the level of $p < .05$. Further, the tests using all items had effect sizes ranging from .00 to .15 with a mean of .07. The tests using the common items had effect sizes ranging from .01 to .09 with a mean of .03. Note BC was the subtest with the largest effect size when including all items ($d = .15$), but was nearly the smallest ($d = .02$) using the common items. This theme of BC showing the largest differences between Forms T1 and T2 does reappear. Although some of the subtests showed significant differences, this is attributable to the large sample size and was expected – the significance does not necessarily indicate practically meaningful inequality between the two forms. As summarized in Appendices O (all items) and P (common items), similar results were observed across the stratified samples.

Table 19. Unstratified Sample Between-Subj. Alt. Forms Reliability – All Items

Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES	Abs. M Diff.
VA	14.74	4.37	14.58	14.74	.03**	.15
AR	13.33	4.99	13.17	13.33	.03**	.16
WK	14.86	5.59	14.52	14.86	.06**	.34
MK	14.18	5.34	14.18	14.18	.00	.00
RC	17.07	4.02	17.51	17.07	.11**	.44
PS	10.54	3.96	11.03	10.54	.12**	.48
TR	27.39	6.29	26.84	27.39	.09**	.55
IC	14.62	6.76	14.57	14.62	.01	.05
BC	15.24	5.95	16.09	15.24	.15**	.84
AI	9.43	4.32	8.91	9.43	.13**	.52

** $p \leq .010$; unmarked are non-significant p .

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; ES = Effect Size = Cohen's d ; Abs. M Diff. = Absolute Mean Difference. All above results use Welch's t -test.

Total $N_{T1} = 32,952$. Total $N_{T2} = 32,952$. All t -tests were Welch's t -tests. For subtest-specific N , please refer to the Attachment 2.

Table 20. Unstratified Sample Between-Subj. Alt. Forms Reliability – Common Items

Subtest	# of Common Items	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES	Abs. M Diff.
VA	8	5.03	1.76	4.88	1.77	.09**	.15
AR	8	3.83	1.92	3.86	1.95	.02*	.04
WK	9	4.88	2.36	4.90	2.35	.01	.02
MK	9	4.83	2.23	4.79	2.15	.02*	.04
RC	0	NA	NA	NA	NA	NA	NA
PS	7	3.83	1.77	3.93	1.77	.06**	.11
TR	10	7.13	1.88	7.07	1.87	.04**	.07
IC	13	6.17	3.81	6.08	3.86	.02**	.09
BC	15	8.57	3.30	8.63	3.27	.02*	.05
AI	7	3.25	1.66	3.29	1.64	.03**	.04

* $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; ES = Effect Size = Cohen's d ; Abs. M Diff. = Absolute Mean Difference; NA = Not Available because RC had no common items. Total N_{T1} = 32,952. Total N_{T2} = 32,952. For subtest-specific N , please refer to the Attachment 2.

3.5.3 Within-Subjects Alternate Forms Reliability Results

Table 21 displays the results of the within-subjects alternate forms reliability analyses across all items. Effect sizes were small and absolute mean differences were trivial across the stratified and unstratified samples. Two tests (MK and IC) showed non-significant differences while two others (AR and RC) showed significance only at the level of $p < .05$. The effect sizes ranged from $d = .04$ to $.22$, with a mean of $.12$. As mentioned before, the largest effect size ($d = .22$) and the largest absolute mean difference (1.01) was found for BC. We also examined the correlations between scores on Forms T1 and T2. Since the order of testing was balanced, we expected to see strong correlations. Consistent with our expectations, the correlations ranged from $.52$ (TR) to $.78$ (WK) with a mean of $.65$ ($p \leq .01$). This provides sufficient evidence for the fact that T1 and T2 are parallel. Similar results were observed across the stratified samples (see Appendix Q).

Table 21. Unstratified Sample Within-Subj. Alt. Forms Reliability – All Items

Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	r_{T1-T2}	ES	Abs. M Diff.
VA	12.31	4.38	12.47	4.12	.59**	.04**	.16
AR	10.50	4.48	10.62	4.26	.68**	.03*	.12
WK	11.42	5.07	11.58	5.26	.78**	.05**	.16
MK	11.81	4.47	11.75	4.63	.68**	.02	.06
RC	15.07	4.47	14.96	4.10	.62**	.03*	.11
PS	9.29	3.42	8.83	3.44	.69**	.17**	.45
TR	25.34	6.67	25.88	6.40	.52**	.08**	.54
IC	13.14	6.38	13.06	6.76	.68**	.02	.08
BC	14.72	5.10	13.70	5.57	.63**	.22**	1.01
AI	8.09	3.81	8.63	4.27	.67**	.17**	.54

* $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; r = Correlation; ES = Effect Size = Cohen's d ; Abs. M Diff. = Absolute Mean Difference. Total N_{T1} = 4,982. Total N_{T2} = 4,982. For subtest-specific N , please refer to the Attachment 2.

3.5.4 LMEM Results

Table 22 displays the output of the LMEM analysis. Consistent with the findings obtained from the alternate forms reliability analyses, the results showed that there was a non-significant interaction between ‘Order’ and ‘Form’ for eight of the 10 subtests. Although the interaction terms for VA ($F(5,277) = 5.71, p \leq .05$) and IC ($F(5,275) = 8.42, p \leq .01$) reached significance, their 90% Confidence Intervals (CI’s) around the partial omega squared (a measure of an effect size; ω^2) included zero. This suggests that there is no practical impact of the interaction between Form and Order. Since there is no significant interaction, we can interpret the main effects. The main effects of ‘Order’ were all statistically significant. PS was the only subtest for which the partial omega squared was null with CI including zero. The results of the main effects of Order will be further discussed in Section 3.6. Eight of 10 main effects of ‘Form’ were statistically significant. Of these eight main effects, the CI for the partial omega squared included zeros for five subtests. The three subtests for which the CI did not include zero were PS ($\omega^2 = .03$, CI [.01; .03]), BC ($\omega^2 = .06$, CI [.05; .07]), and AI ($\omega^2 = .03$, CI [.02; .04]). Two subtests (MK and IC) did not reach significance. All effect sizes were null to small. Interestingly, the LMEM results confirmed the previous findings that the BC subtest was the least equivalent in comparison to the rest of the subtests.

Table 22. Unstratified Sample Alt. Forms Reliability – LMEM

Subtest	Effects	<i>F</i>	<i>Sig.</i>	ω^2	CI Lower	CI Upper
VA	ME: Order	1085.35	.00	.17	.16	.19
	ME: Form	13.50	.00	.00	.00	.01
	INT: Order x Form	5.71	.02	.00	.00	.00
AR	ME: Order	488.44	.00	.08	.07	.10
	ME: Form	7.80	.01	.00	.00	.00
	INT: Order x Form	.11	.74	.00	.00	.00
WK	ME: Order	614.53	.00	.10	.09	.12
	ME: Form	13.18	.00	.00	.00	.00
	INT: Order x Form	.18	.67	.00	.00	.00
MK	ME: Order	644.59	.00	.11	.10	.12
	ME: Form	1.44	.23	.00	.00	.00
	INT: Order x Form	.01	.94	.00	.00	.00
RC	ME: Order	440.62	.00	.08	.07	.09
	ME: Form	8.98	.00	.00	.00	.00
	INT: Order x Form	3.42	.06	.00	.00	.00
PS	ME: Order	9.87	.00	.00	.00	.00
	ME: Form	143.49	.00	.03	.02	.03
	INT: Order x Form	.13	.72	.00	.00	.00
TR	ME: Order	874.49	.00	.14	.13	.16
	ME: Form	38.17	.00	.01	.00	.01
	INT: Order x Form	1.32	.25	.00	.00	.00
IC	ME: Order	1306.45	.00	.20	.18	.21
	ME: Form	2.28	.13	.00	.00	.00
	INT: Order x Form	8.42	.00	.00	.00	.00
BC	ME: Order	1273.49	.00	.19	.18	.21
	ME: Form	327.86	.00	.06	.05	.07
	INT: Order x Form	.41	.52	.00	.00	.00
AI	ME: Order	625.53	.00	.11	.09	.12
	ME: Form	160.29	.00	.03	.02	.04
	INT: Order x Form	.01	.94	.00	.00	.00

Note. ME = Main Effect; INT = Interaction Term; *F* = *F*-Statistic; *Sig.* = Significance Level; ω^2 = Omega Squared Effect Size; CI = Confidence Interval; CI Lower = Lower Boundary of Confidence Interval; CI Upper = Upper Boundary of Confidence Interval. Total *N* = 10,558. For subtest-specific *N*, please refer to the Attachment 2.

Based on the interaction terms, on the main effects of Form, and on the BC subtest findings, it seems that (1) Forms T1 and T2 do not need to be equated; but (2) BC should be revised at the item-level in the next iterations of the AFOQT to eliminate problematic (non-equivalent) items.

3.6 RQ6: Who Retests? Do Retesters Perform Better When They Retest?

To determine whether or not there were quantitative and qualitative differences among the test-takers who choose to retest, additional analyses were performed.

3.6.1 First Attempt Scores for Retesters vs. Non-Retesters

In this section, we address the question: Is there a mean score difference between those test-takers

who chose to retest (referred to as ‘*retesters*’) versus those who chose not to retest (referred to as ‘*non-retesters*’)? We hypothesized that retesters would score lower on their first attempt compared to non-retesters.

Table 23 contains results of the independent samples *t*-test analyses that were run to address the aforementioned hypothesis. As expected, retesters scored significantly lower on their first attempt compared to non-retesters. Effect sizes ranged from $d = .35$ (AI) to $d = .75$ (VA), with retesters scoring lower on average than non-retesters by 2.88 points.

Table 23. Unstratified Sample First Attempt Scores for Non-Retesters vs. Retesters

Subtest	$M_{\text{Non-Retesters}}$	$SD_{\text{Non-Retesters}}$	$M_{\text{Retesters}}$	$SD_{\text{Retesters}}$	t^{**}	df	ES	M Diff.
VA	14.92	4.39	11.63	4.15	52.59	69046	.75	3.29
AR	13.52	5.05	10.08	4.33	47.91	69044	.69	3.43
WK	15.00	5.46	10.96	5.08	51.90	69034	.74	4.04
MK	14.43	5.20	11.21	4.57	43.53	69034	.62	3.21
RC	17.51	4.07	14.51	4.31	51.12	69025	.73	3.00
PS	10.92	3.94	9.01	3.42	34.15	68995	.49	1.91
TR	27.34	6.36	24.43	6.16	31.97	68987	.46	2.91
IC	14.81	6.63	11.92	6.26	30.43	68971	.44	2.88
BC	15.84	5.66	13.17	5.18	33.13	68990	.48	2.67
AI	9.29	4.17	7.84	3.73	24.40	68891	.35	1.45

**Significant at $p \leq .010$.

Note. $M_{\text{Non-Retesters}}$ = Mean of Non-Retesters; $SD_{\text{Non-Retesters}}$ = Standard Deviation of Non-Retesters; $M_{\text{Retesters}}$ = Mean of Retesters; $SD_{\text{Retesters}}$ = Standard Deviation of Retesters; t = *t*-statistic; df = Degrees of Freedom; ES = Effect Size = Cohen’s *d*; M Diff. = Mean Difference. Total $N_{\text{Non-Retesters}}$ = 63,772. Total $N_{\text{Retesters}}$ = 5,279. For subtest-specific *N*, please refer to the Attachment 2.

3.6.2 Demographic Differences between Retesters and Non-Retesters

In this section, we address the question: Are there demographic differences between retesters and non-retesters? We conducted frequency analyses and independent sample *t*-tests where possible (i.e., on continuous variables including Age and Education Level) on all demographic variables. Table 24 displays the results of the independent samples *t*-tests comparing the non-retesters and retesters based on age and education. Table 25 shows the results of the demographics frequencies analyses. We found that compared to non-retesters, retesters generally: (1) are younger ($M_{\text{Non-Retesters}}$ = 23.82 years old; $M_{\text{Retesters}}$ = 22.21 years old); (2) are more likely to be female (Non-Retesters: 26%; Retesters: 31%); (3) are more likely to be representatives of a racial and ethnic subgroup (most prevalent for Black test-takers; non-retesters: 13%; retesters: 21%); (4) have a lower academic degree; and (5) are from ROTC (non-retesters: 34%; retesters: 60%). We did not observe large differences for test-takers from different socio-economic statuses.

Table 24. Unstratified Sample Non-Retesters vs. Retesters: Age and Education

Variable	$M_{\text{Non-Retesters}}$	$SD_{\text{Non-Retesters}}$	$M_{\text{Retesters}}$	$SD_{\text{Retesters}}$	t^{**}	df	ES	M Diff.
Age	23.82	5.10	22.21	4.52	22.20	68591	.32	1.61
Education Level	14.91	1.78	14.20	1.61	27.90	68924	.40	.71

**Significant at $p \leq .010$.

Note. $M_{\text{Non-Retesters}}$ = Mean of Non-Retesters; $SD_{\text{Non-Retesters}}$ = Standard Deviation of Non-Retesters; $M_{\text{Retesters}}$ = Mean of Retesters; $SD_{\text{Retesters}}$ = Standard Deviation of Retesters; t = *t*-statistic; df = Degrees of Freedom; ES = Effect Size = Cohen’s *d*; M Diff. = Mean Difference. Total $N_{\text{Non-Retesters}}$ = 63,772. Total $N_{\text{Retesters}}$ = 5,279.

Table 25. Unstratified Sample Non-Retesters vs. Retesters: Demographic Frequencies

Sample Characteristic	Non-Retesters		Retesters		Absolute % Diff.
	N	%	N	%	
Sex					
Male	47,226	74%	3,626	69%	5%
Female	16,468	26%	1,647	31%	5%
Unknown	78	0%	6	0%	0%
Race*					
White	49,598	78%	3,506	66%	11%
Non-White	11,629	18%	1,439	27%	9%
Unknown	2,545	4%	334	6%	2%
Non-Black/African American	51,466	81%	3,838	73%	8%
Black/African American	8,495	13%	1,087	21%	7%
Unknown	3,811	6%	354	7%	1%
Non-Asian	52,680	83%	4,177	79%	3%
Asian	6,423	10%	616	12%	2%
Unknown	4,669	7%	486	9%	2%
Non-AI/AN	55,316	87%	4,403	83%	3%
AI/AN	3,416	5%	335	6%	1%
NA	5,040	8%	541	10%	2%
Non-NHPI	57,066	89%	4,567	87%	3%
NHPI	1,575	2%	168	3%	1%
NA	5,131	8%	544	10%	2%
Ethnicity					
Non-Hispanic	53,639	84%	4,077	77%	6%
Hispanic	8,898	14%	1,059	20%	7%
Unknown	1,235	2%	143	3%	1%
Socio-Economic Status					
Much higher than average	3,218	5%	239	5%	1%
Somewhat higher than average	18,244	29%	1,374	26%	3%
Average	24,500	38%	2,122	40%	2%
Somewhat lower than average	10,896	17%	914	17%	0%
Much lower than average	3,749	6%	373	7%	1%
Declined to respond	2,969	5%	236	4%	0%
Unknown	196	0%	21	0%	0%
Years of Education Completed					
12	2143	3%	285	5%	2%
13	18,096	28%	2,280	43%	15%
14	8,315	13%	846	16%	3%
15	6,699	11%	532	10%	0%
16	19,325	30%	906	17%	13%
17	4,157	7%	205	4%	3%
18	3,400	5%	146	3%	3%
19	727	1%	31	1%	1%
20	404	1%	18	0%	0%
21+	396	1%	15	0%	0%
Unknown	110	0%	15	0%	0%
Highest Academic Degree					
High School Diploma	30,297	48%	3,482	66%	18%
Associate's Degree	6,537	10%	597	11%	1%
Bachelor's Degree	23,393	37%	1,032	20%	17%
Master's Degree	3,182	5%	138	3%	2%
Doctoral Degree	212	0%	5	0%	0%
Unknown	151	0%	25	0%	0%
Accession Source					
OTS-CIV	14,892	23%	418	8%	15%
OTS-AD	9,629	15%	644	12%	3%
AACP	332	1%	6	0%	0%
USAFA	7,615	12%	438	8%	4%
ROTC	21,993	34%	3,185	60%	26%
ANG	6,779	11%	424	8%	3%
AFRES	2,532	4%	164	3%	1%

Note. N = Sample Size; Absolute % Diff. = Absolute Percent Difference.

3.6.3 Effects of Retesting on Scores

In this section, we address the questions: (1) Did retesters improve their scores upon retesting?; (2) If they did, did they improve more on the parallel forms or the same form?; and (3) Which subtests saw the highest/lowest score improvements? We performed several analyses. In terms of our methodology, we used three different samples. The first sample consisted of test-takers who tested on T1 twice. The second sample consisted of test-takers who tested on T2 twice. These two samples were used previously in Section 3.4.2. Specifically, they were used to examine test-retest reliability (the correlation between the testing attempts on the same form). The analyses in this section are different because they assess how much, if at all, people improved when they retested on the same form using paired-samples *t*-tests. The third sample consisted of test-takers who tested on Form T1 and then on T2 or on Form T2 and then on T1. This sample was also used previously for the within-subjects alternate forms reliability analyses described in Section 3.5. However, instead of examining the means of the two test forms and testing the differences between test forms, the analyses in this section examine the means of earlier and later test attempts, and test the differences between the testing attempts.

We tested the mean score difference between earlier and later attempts, using the paired-samples *t*-tests. Table 26 summarizes the retesting effects on the same form and Table 27 on the parallel forms. Those who retested on the same form (T1 then T1 or T2 then T2) moderately improved their scores. Specifically, on T1, scores improved as little as .18 points (PS) and as much as 1.91 points (IC), with an average improvement of 1.12 points. On T2, scores improved as little as .18 points (PS) and as much as 1.73 points (VA), with an average improvement of 1.06 points. As mentioned before, the AFOQT policy allows up to three retests on parallel forms (Fedrigo, 2021). With only two parallel forms, it is possible that the test-takers would be exposed to the same form twice (e.g., T1 then T2 then T1 again). However, the minimum interval between test attempts is 150 days. Thus, the minimum interval between the same forms should be roughly 300 days. With almost a year between attempts of the same form, recall and practice effects should be minimal. Appendix R contains the stratified samples retesting effects on Form T1. Appendix S contains the stratified samples retesting effects on Form T2.

For those who retested on the parallel forms (T1 then T2 or T2 then T1), scores improved moderately. Specifically, the improvement was as little .12 points (PS) and as much as 2.39 points (TR), with an average improvement of 1.39 points. Appendix T contains the stratified samples retesting effects on the parallel forms.

The results for the stratified samples across the same form were similar in that the retesting scores tended to improve, with ROTC test-takers improving their scores the most (on average of 1.50 points on Form T1 and 1.37 points on T2). The results for the parallel forms were more optimistic in that there was a higher increase in retesting scores (about 1 point increase across all subtests and accession sources).

The aforementioned findings are consistent with the LMEM analyses, which revealed significant main effects of 'Order.' This indicates that order did statistically and practically drive the variance in the test scores. The exception was with PS, in which people were not able to improve their scores between testing attempts, which could be due to the fact that since PS is not included in the

operational composites there may be a lack of motivation to try to improve performance (see Table 22).

This pattern of findings suggests that: (1) retesters improve their scores upon retesting; (2) the improvements are similar on the same and parallel forms (average effect size for T1-T1 was $d = .36$; average effect size for T2-T2 was $d = .34$, and average effect size for the parallel forms was $d = .35$); (3) the highest score improvements were observed for VA, IC, and TR and the lowest score improvements were observed for PS.

Table 26. Unstratified Sample Retesting Effects on the Same Form

Subtest	T1-T1							
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES	Abs. M Diff.
VA	12.24	4.24	13.65	3.89	-14.10**	713	.53	1.41
AR	10.95	4.52	11.77	4.48	-8.06**	713	.30	.82
WK	11.34	5.26	12.38	5.23	-10.03**	712	.38	1.04
MK	11.86	4.86	12.95	4.60	-10.22**	711	.38	1.09
RC	14.85	4.26	15.92	4.05	-10.46**	712	.39	1.07
PS	9.09	3.51	9.28	3.54	-2.44*	707	.09	.18
TR	26.17	6.40	27.63	6.60	-8.14**	708	.31	1.46
IC	12.63	6.85	14.53	6.96	-12.28**	706	.46	1.91
BC	13.72	5.77	15.14	5.96	-10.21**	708	.38	1.42
AI	8.47	4.41	9.26	4.59	-9.54**	705	.36	.79
Subtest	T2-T2							
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES	Abs. M Diff.
VA	11.65	4.59	13.38	4.13	-13.47**	563	.57	1.73
AR	9.91	4.45	10.76	4.39	-7.20**	563	.30	.84
WK	11.31	5.41	12.21	5.23	-7.87**	563	.33	.90
MK	10.97	4.29	12.05	3.96	-8.95**	563	.38	1.08
RC	14.66	4.70	15.78	4.44	-8.92**	563	.38	1.13
PS	9.01	3.32	9.19	3.36	-2.18*	562	.09	.18
TR	24.49	6.65	25.99	7.20	-6.26**	560	.26	1.50
IC	12.14	6.30	13.41	6.54	-8.22**	562	.35	1.27
BC	14.21	5.07	15.70	5.12	-11.24**	561	.47	1.50
AI	7.64	3.77	8.12	4.07	-5.18**	559	.22	.48

* $p \leq .050$; ** $p \leq .010$.

Note. M_{Earlier} = Mean of Earlier Test Administration; SD_{Earlier} = Standard Deviation of Earlier Test Administration; M_{Later} = Mean of Later Test Administration; SD_{Later} = Standard Deviation of Later Test Administration; t = t -test; df = Degrees of Freedom; ES = Effect Size = Cohen's d ; Abs. M Diff. = Absolute Mean Difference. Total $N_{\text{T1-T1}}$ = 714. Total $N_{\text{T2-T2}}$ = 564. For subtest-specific N , please refer to the Attachment 2.

Table 27. Unstratified Sample Retesting Effects on the Parallel Forms

Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t^{**}	df	ES	Abs. M Diff.
VA	11.59	4.17	13.19	4.18	-32.37	4980	.46	1.60
AR	10.04	4.31	11.08	4.37	-21.84	4980	.31	1.04
WK	10.94	5.08	12.06	5.19	-24.26	4974	.34	1.12
MK	11.18	4.55	12.38	4.47	-24.73	4976	.35	1.20
RC	14.51	4.33	15.52	4.20	-19.63	4973	.28	1.01
PS	9.00	3.42	9.12	3.45	-3.12	4960	.04	.12
TR	24.42	6.16	26.81	6.69	-28.12	4959	.40	2.39
IC	11.93	6.26	14.26	6.67	-35.03	4951	.50	2.33
BC	13.20	5.17	15.22	5.36	-33.35	4957	.47	2.02
AI	7.83	3.74	8.90	4.28	-23.90	4945	.34	1.07

**Significant at $p \leq .010$.

Note. M_{Earlier} = Mean of the Earlier Test Administration; SD_{Earlier} = Standard Deviation of the Earlier Test Administration; M_{Later} = Mean of the Later Test Administration; SD_{Later} = Standard Deviation of the Later Test Administration; t = t -test; df = Degrees of Freedom; ES = Effect Size = Cohen's d ; Abs. M Diff. = Absolute Mean Difference. Total N_{Earlier} = 4,982. Total N_{Later} = 4,982. For subtest-specific N , please refer to the Attachment 2.

3.6.4 Score Improvements between Multiple Retests

In this section, we address the question: To what extent did scores improve between multiple retesting attempts? When considering the AFOQT retesting policy (Fedrigo, 2021), those who test twice on T1 or T2 should have taken the AFOQT a total of three times. Then, when comparing their first and third attempts (T1/T1 or T2/T2), smaller improvements are observed than on their first and second attempts (T1/T2 or T2/T1). To directly explore if people improve less on their third attempts, we examined mean differences of those who took the test twice total and their test scores, and the mean differences of people who took the test three times total and their test scores.

Table 28 displays the results of these analyses. Test-takers who take the test twice have similar total score improvement in comparison to those who take the test three times. However, those who take the test three times have lower first time attempt scores than those who only take the test twice. It would seem that although allowing a third attempt does enable test takers to further improve their scores, they still do not outperform those who only needed to attempt twice. And that being said, those who need to retest, even once, do not outperform those who choose not to retest at all.

Table 28. Unstratified Sample Two and Three Attempt Score Increases

Subtest	T1			T2		
	Attempt or Increase	1x Retesters	2x Retesters	Attempt or Increase	1x Retesters	2x Retesters
VA	1 st Attempt	12.26	10.15	1 st Attempt	11.85	9.59
	2 nd Attempt	13.58	11.30	2 nd Attempt	13.63	10.45
	3 rd Attempt		12.47	3 rd Attempt		12.64
	Increase 1 st to 2 nd	1.32	1.15	Increase 1 st to 2 nd	1.78	.86
	Increase 2 nd to 3 rd		1.17	Increase 2 nd to 3 rd		2.19
	Total Increase	1.32	2.32	Total Increase	1.78	3.05
AR	1 st Attempt	10.61	8.93	1 st Attempt	10.32	8.45
	2 nd Attempt	11.65	9.16	2 nd Attempt	11.37	8.91
	3 rd Attempt		9.96	3 rd Attempt		9.83
	Increase 1 st to 2 nd	1.04	.23	Increase 1 st to 2 nd	1.05	.46
	Increase 2 nd to 3 rd		.80	Increase 2 nd to 3 rd		.92
	Total Increase	1.04	1.03	Total Increase	1.05	1.38
WK	1 st Attempt	11.60	8.79	1 st Attempt	11.37	8.92
	2 nd Attempt	12.70	9.39	2 nd Attempt	12.52	8.86
	3 rd Attempt		10.21	3 rd Attempt		10.37
	Increase 1 st to 2 nd	1.10	.60	Increase 1 st to 2 nd	1.15	-.06
	Increase 2 nd to 3 rd		.82	Increase 2 nd to 3 rd		1.51
	Total Increase	1.1	1.42	Total Increase	1.15	1.45
MK	1 st Attempt	11.55	10.18	1 st Attempt	11.44	9.98
	2 nd Attempt	12.81	10.57	2 nd Attempt	12.68	10.23
	3 rd Attempt		11.46	3 rd Attempt		11.70
	Increase 1 st to 2 nd	1.26	.39	Increase 1 st to 2 nd	1.24	.25
	Increase 2 nd to 3 rd		.89	Increase 2 nd to 3 rd		1.47
	Total Increase	1.26	1.28	Total Increase	1.24	1.72
RC	1 st Attempt	14.91	13.00	1 st Attempt	14.95	12.58
	2 nd Attempt	15.83	13.30	2 nd Attempt	16.15	12.94
	3 rd Attempt		14.83	3 rd Attempt		14.76
	Increase 1 st to 2 nd	.92	.30	Increase 1 st to 2 nd	1.20	.36
	Increase 2 nd to 3 rd		1.53	Increase 2 nd to 3 rd		1.82
	Total Increase	.92	1.83	Total Increase	1.20	2.18
PS	1 st Attempt	9.11	7.68	1 st Attempt	9.46	7.89
	2 nd Attempt	9.23	7.84	2 nd Attempt	9.59	7.97
	3 rd Attempt		7.85	3 rd Attempt		8.41
	Increase 1 st to 2 nd	.12	.16	Increase 1 st to 2 nd	.13	.08
	Increase 2 nd to 3 rd		.01	Increase 2 nd to 3 rd		.44
	Total Increase	.12	.17	Total Increase	.13	.52
TR	1 st Attempt	25.07	24.85	1 st Attempt	23.31	23.51
	2 nd Attempt	27.26	25.87	2 nd Attempt	26.78	25.12
	3 rd Attempt		27.06	3 rd Attempt		26
	Increase 1 st to 2 nd	2.19	1.02	Increase 1 st to 2 nd	3.47	1.61
	Increase 2 nd to 3 rd		1.19	Increase 2 nd to 3 rd		.88
	Total Increase	2.19	2.21	Total Increase	3.47	2.49

IC	1 st Attempt	12.55	10.34	1 st Attempt	12.02	10.64
	2 nd Attempt	14.36	11.87	2 nd Attempt	14.68	12.65
	3 rd Attempt		13.77	3 rd Attempt		13.01
	Increase 1 st to 2 nd	1.81	1.53	Increase 1 st to 2 nd	2.66	2.01
	Increase 2 nd to 3 rd		1.90	Increase 2 nd to 3 rd		.36
	Total Increase	1.81	3.43	Total Increase	2.66	2.37
BC	1 st Attempt	13.14	11.83	1 st Attempt	13.97	12.83
	2 nd Attempt	14.99	12.87	2 nd Attempt	16.02	14.03
	3 rd Attempt		14.55	3 rd Attempt		15.14
	Increase 1 st to 2 nd	1.85	1.04	Increase 1 st to 2 nd	2.05	1.2
	Increase 2 nd to 3 rd		1.68	Increase 2 nd to 3 rd		1.11
	Total Increase	1.85	2.72	Total Increase	2.05	2.31
AI	1 st Attempt	8.35	7.25	1 st Attempt	7.72	6.85
	2 nd Attempt	9.47	7.41	2 nd Attempt	8.77	7.45
	3 rd Attempt		8.26	3 rd Attempt		7.55
	Increase 1 st to 2 nd	1.12	.16	Increase 1 st to 2 nd	1.05	.60
	Increase 2 nd to 3 rd		.85	Increase 2 nd to 3 rd		.10
	Total Increase	1.12	1.01	Total Increase	1.05	.70

Few previous researchers have examined effects of retesting on the AFOQT. Arth (1986) explored test-retest reliabilities for AFOQT Form O and found reliabilities for composites to be between .78 and .88. Additionally, the author observed that those who chose to retest had lower first attempt scores than those who did not choose to retest. To follow this review, Carretta and Ree (1997) also examined the effects of retesting for AFOQT Forms O, P1, and P2 and found that (1) those who chose to retest scored lower on their first testing attempt, (2) there were moderate to large increases in performance on retests, and (3) the amount of increase was a function of test content. Specifically, the authors found the greatest improvement on IC, and the least improvement on AR, MK, and GS. Our findings support the work performed previously.

In sum, our findings seem to indicate that: (1) retesters score lower on their first attempt compared to non-retesters; (2) there are demographic differences between retesters and non-retesters; (3) retesters tend to improve their scores whether it be on the same form or parallel forms; and (4) retesters tend to improve their scores with every retest.

3.7 RQ7: Was There a Test Security Breach?

Test security is always a concern especially for enterprise-wide high-stakes employment tests, such as the AFOQT. Although proactive steps have been taken to ensure test security (e.g., creation of alternate forms), it is prudent to continue scrutinizing the test regardless of test media (i.e. paper-and-pencil format or electronic format) for potential indicators of test security breach. Statistically, we tested for breach of test security by examining whether or not there is a change in test parameters over time. To do that, the between-subject mean scores for the first-time test-takers who took the AFOQT between January 2015 and August 2016 (earlier group) were compared against the mean scores for the first-time test-takers who took the AFOQT between May 2018 and December 2019 (later group). Note that similar to Morath et al. (2008) methodology, both groups

were of equal size within each stratification and were randomly drawn from a larger group. For example, the total unstratified sample size for early group on Form T1 was 9,114 and for later group was 10,810. With the goal to maximize the usable sample size, 9,000 random cases were selected for the stability analysis. Same process was repeated for the accession sources.

Table 29 displays the means, standard deviations, effect sizes, and absolute mean differences for the unstratified sample. The comparison of the means and standard deviations between and within the parallel forms suggests that some mean scores increased (e.g., T1: MK by .11 points; T2: MK by .19 points) and some decreased (e.g., T1: VA by -.53 points; T2: VA by -.29 points). The largest mean score difference was observed for TR, with a decrease of -.77 points on T1 and of -.86 points on T2. Although some of the mean differences were statistically significant, they were not practically significant due to large sample sizes. Furthermore, the comparison of the effect sizes between and within the parallel forms revealed that they were trivial with Cohen's *d* less than .40. Overall, these results suggest that a past security breach is unlikely. Appendix U contains the stability statistics stratified by accession source. The results across the accession sources were similar to the results discussed above.

Table 29. Unstratified Sample Stability Analysis

Subtest	T1				ES	Abs. <i>M</i> Diff.
	<i>M</i> _{Earlier}	<i>SD</i> _{Earlier}	<i>M</i> _{Later}	<i>SD</i> _{Later}		
VA	14.98	4.41	14.45	4.36	.12**	.53
AR	13.52	5.04	13.09	4.95	.09**	.43
WK	15.18	5.66	14.53	5.56	.12**	.65
MK	14.06	5.38	14.17	5.32	.02	.11
RC	17.1	4.04	16.95	4.05	.04*	.14
PS	10.36	4.02	10.6	3.96	.06**	.24
TR	27.82	6.36	27.05	6.25	.12**	.77
IC	14.76	6.81	14.32	6.79	.06**	.44
BC	15.51	6.02	15.00	5.96	.09**	.51
AI	9.46	4.31	9.36	4.28	.02	.10
Subtest	T2				ES	Abs. <i>M</i> Diff.
	<i>M</i> _{Earlier}	<i>SD</i> _{Earlier}	<i>M</i> _{Later}	<i>SD</i> _{Later}		
VA	14.7	4.58	14.41	4.51	.06**	.29
AR	13.24	5.23	13.02	5.1	.04**	.22
WK	14.66	5.61	14.38	5.35	.05**	.27
MK	14	5.16	14.2	5.04	.04*	.19
RC	17.51	4.35	17.44	4.33	.02	.07
PS	10.88	3.89	11.03	3.88	.04*	.15
TR	27.39	6.53	26.52	6.37	.13**	.86
IC	14.55	6.51	14.45	6.49	.02	.10
BC	16.34	5.37	15.92	5.27	.08**	.42
AI	8.84	3.94	8.88	3.91	.01	.04

* $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. *M*_{Earlier} = Mean of Earlier Group; *SD*_{Earlier} = Standard Deviation of Earlier Group; *M*_{Later} = Mean of Later Group; *SD*_{Later} = Standard Deviation of Later Group; ES = Effect Size = Cohen's *d*; Abs. *M* Diff. = Absolute Mean Difference. Total T1 *N* = 9,000. Total T2 *N* = 8,000. For subtest-specific *N*, please refer to the Attachment 2.

Reviewing the results of the stability analysis for the stratified samples between and within the parallel forms revealed that most of the accession sources have small effect sizes (Cohen's $d \leq .40$), with some subtests increasing, some decreasing, and some staying almost the same. Note that the sample size for AECF was too small to calculate the stability parameters (T1: $N_{\text{Earlier}} = 54$, $N_{\text{Later}} = 26$; T2: $N_{\text{Earlier}} = 51$, $N_{\text{Later}} = 34$). One exception to this pattern of results was USAFA, for which one subtest, namely AI, was found to have moderate effect size (T1: Cohen's $d = .44$; T2: Cohen's $d = .42$). Curiously, the means across most subtests (except T2 MK) decreased between the two administration periods, with the largest decrease of 1.70 on T1 AI and of 1.47 on T2 AI. There are a couple of possible explanation to this counter-intuitive trend. First, USAFA had trouble meeting its quotas for the rated career fields so they "encouraged" the cadets to take the AFOQT. Some of these cadets did not want a rated assignment so they intentionally did poorly on the test. Second, USAFA cadets have a common understanding that their AFOQT performance does not receive much weight by the selection boards so did not put forth their best effort (low motivation).

Since the previous Form T technical reports do not report the stability results, we compared our findings against those from Morath et al. (2008) for the AFOQT Form S with the aforementioned limitations. For the most part, the pattern of results was similar with some subtest mean score increasing and others decreasing. While some mean score differences were statistically significant due to large sample sizes, they were deemed as practically trivial. The conclusion was that a past test security breach is unlikely.

4.0 DISCUSSION

The purpose of these analyses was to psychometrically evaluate the 10 AFOQT Form T cognitive subtests. The text below summarize our findings and provides recommendations for how to improve the measurement of cognitive ability in future iterations of the AFOQT. Table 30 summarizes the findings of the current research, color coded based on the current settings of the AFOQT (i.e., select out, operational composites do not include non-cognitive subtests).

Table 30. Subtest Summary

Subtest	Difficulty	Discriminability	Subgroup Differences	Factor Loadings
VA	✓	✗	✓	g
AR	✓	✓	✓	g
WK	✓	✓	✓	Verbal Ability
MK	✓	✓	✓	g
RC	✗	✗	✓	g
PS	✓	✓	✓	g
TR	✗	✓	✓	Perceptual Speed
IC	✓	✓	✗	g & Aviation Knowledge
BC	✓	✓	✓	Spatial Ability
AI	✓	✓	✗	g

Note. ✓ - Acceptable; ✓ - Marginally Acceptable; ✗ - Unacceptable.

4.1 Summary of Psychometric Properties

As Section 3.1 showed, most Form T cognitive subtests showed acceptable psychometric properties in terms of the measures of central tendency, difficulty, and discriminability both within and between the parallel forms (T1 and T2). With that said, RC and TR emerged as the least difficult subtests and AI as the most difficult subtest. VA and RC had rather poor discriminability, while IC had high discriminability. TR, IC, and BC showed the highest evidence for speededness. BC also showed higher evidence for overt guessing. Finally, there is initial evidence to suggest that RGL may impact the subtest difficulty. Based on these results, we recommend that the subtests be designed (1) to target lower ranges of cognitive ability (around the cutoff scores); (2) to target moderate to high levels of item discriminability; and (3) with appropriate RGL.

4.2 Summary of Subtest Inter-Correlations

All Form T cognitive subtests are moderately related to each other and are cognitively loaded, as discussed in Section 3.2. The multivariate correlations were higher for the academic achievement subtests (VA, AR, MK, WK, and RC), included in the first half of the AFOQT, and lower for specialized knowledge subtests (PS, TR, IC, BC, and AI), included in the second half of the AFOQT. Stronger correlations were observed among theoretically linked subtests, such as verbal ability subtests (e.g., VA and WK). Weaker correlations were observed for less theoretically linked subtests, such as perceptual speed and verbal ability subtests (TR and VA). The highest correlations were found between AR and MK and between VA and RC suggesting potential redundancy in measurement. Consistent with the literature, the bi-factor factor model emerged as the best fitting in terms of the model-data fit indices. It supports the existence of a general cognitive ability factor (*g*) and of five content factors (Verbal Ability, Quantitative Ability, Spatial Ability, Aviation Knowledge, and Perceptual Speed) that explain subtest score variance. With that said, VA, AR, MK, RC, and AI were found to load mostly on *g*; whereas, WK, TR, and BC were found to load mostly on their respective content factors. Based on these results and on a general literature findings, we recommend that the future iterations of the AFOQT include cognitive subtests (1) that are theoretically linked to officership performance criteria and (2) that each measure unique aspects of cognitive ability to avoid redundancy and enhance competency coverage.

4.3 Summary of Mean Score Subgroup Differences

As Section 3.3 revealed, consistent with a broader literature on cognitive ability and with previous technical reports, there were mean score subgroup differences observed on Form T, which may lead to adverse impact. The largest effect sizes were observed for Black test-takers, followed by female test-takers, and then by Asian test-takers, compared to their racial and gender majority subgroup. Additionally, largest effect sizes were observed for ROTC test-takers compared to USAFA test-takers. The subtests with the largest subgroup differences were IC and AI. The subtests with the smallest subgroup differences were VA, MK, and TR. We found evidence for measurement invariance. However, the results of the stepwise regression analyses showed demographic variables explained between 7.3% to 27.9% of the variance in the subtest scores. Based on these findings, we recommend that the future iterations of the AFOQT include (1) cognitive subtests that are relevant to the officership performance criteria but have been found to

result in lower subgroup differences (e.g., fluid intelligence tests); and (2) non-cognitive measures (e.g., personality, Situational Judgment Test).

4.4 Summary of Reliability

As discussed in Section 3.4, all subtests on Form T showed moderate to high internal consistency and test-retest reliability, when retesting on the same form. Recall that according to the AFOQT retesting policy (Fedrigo, 2021), test-takers can test up to three times (a third time with a waiver). Therefore, it is inevitable that those people who test three times will receive the same form at least once. While practice effects can be a concern, the time lag of at least 150 days makes these effects unlikely. The subtest with the highest internal consistency was IC, which may suggest redundancy in measurement. Overall, these findings suggest that the subtests measure the constructs they were designed to measure, even when temporally distributed. Based on these findings, we recommend that (1) the redundant subtests be evaluated for potential item reduction without sacrificing reliability and that (2) test-takers have sufficient time in-between retesting on the same form to eliminate practice effects altogether.

4.5 Summary of Equivalency

As per Section 3.5, all of the Form T subtests showed acceptable alternate forms reliability with sufficient evidence for the equivalency of the parallel forms, which rendered the issue of equating unnecessary. While not needed for Form T, we recommend that the psychometric evaluations of the future AFOQT forms continue to examine the parallel form equivalency in case, if this step is warranted.

4.6 Summary of Retesting

When comparing retesters and non-retesters, the former tended to score lower on their first attempt which subsequently led them to retake the test (Section 3.6). Their scores improved with each attempt but minimally, with the highest improvement observed for TR, IC, and BC and the least improvement observed for PS. Furthermore, we found that those who retest are more likely to be younger, have fewer years of education and/or a lower post-secondary degree, be from the ROTC, be female, and belong to a racial minority.

4.7 Summary of Test Security

Because test security is always a concern, we ran stability analyses (Section 3.7). To do that, the between-subject mean scores for the first-time test-takers who took the AFOQT between January 2015 and August 2016 (earlier group) were compared against the mean scores of those test-takers who took the AFOQT between May 2018 and December 2019 (later group). For the most part, the pattern of results was similar with some subtest mean score increasing and others decreasing. While some mean score differences were statistically significant due to large sample sizes, they were deemed as practically trivial. The conclusion was that a past test security breach is unlikely.

5.0 CONCLUSION

This technical report has detailed the findings from the AFOQT Form T subtest-level analyses project. These results should be utilized to help inform decisions regarding future iterations of the AFOQT. The ultimate decision about the inclusion, exclusion, or revision of the current AFOQT Form T cognitive subtests needs to be based on the holistic interpretation of the data resulting from: Item-level analyses, subtest-level analyses, composite-level analyses, criterion-related validities, adverse impact rates and subgroup qualification rates, critical officer job analysis, and analyses for non-cognitive subtests (SJT, SDI-O).

6.0 REFERENCES

- Aguilar, I. D. (2017). *Comparison of the Air Force Officer Qualifying Test Form T and Form S: Initial item-and subtest-level analyses*. AFCAPS-TR-2017-0002. Randolph AFB, TX: Air Force Personnel Center, Strategic Research and Analysis Branch.
- Aguinis, H., Henle, C. A., & Ostroff, C. (2001). Measurement in work and organizational psychology. *Handbook of industrial, work and organizational psychology, 1*, 27-50
- Allison, P. D. (2001). *Missing data*. Sage publications.
- Arth, T. O. (1986). *Air Force Officer Qualifying Test (AFOQT) retesting effects* (AFHRL-TP-86-8). Brooks AFB, TX. Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Bartholomew, D. J. (2004). *Measuring Intelligence: Facts and fallacies*. Cambridge University Press.
- Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods, 23*(3), 389.
- Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99–136). American Psychological Association.
- Canivez, G. L. (2016). Chapter 12: Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 246-271). Hogrefe Publishers.
- Carretta, T. R., & Ree, M. J. (1997). *The best retest score is the average: findings and implications*, AL/HR-TP-1996-0021. Brooks AFB, TX. Air Force Materiel Command Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Research Division.
- Carretta, T. R., Rose, M. R., & Trent, J. D. (2016). *Air Force Officer Qualifying Test Form T: Initial item-, test-, factor-, and composite-level analyses*, AFRH-RH-WP-TR-2016-0093. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Warfighter Interface Division.
- Carroll, J. B. (1996). A three-stratum theory of intelligence: Spearman's contribution. *Human abilities: Their nature and measurement*, 1996, 1-17.
- Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological*

Bulletin, 112, 155-159.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

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

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

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

Ellis, B. B., & Mead, A. D. (2002). Item analysis: Theory and practice using classical and modern test theory. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 324-33). Blackwell Publishing.

Fedrigio, J. A. (2021, March 26). *Air Force Guidance Memorandum (AFGM) to AFMAN 36 2664*, Personnel Assessment Program [Memorandum]. Department of the Air Force.

Furr, M. R., & Bacharach, V. R. (2008). *Psychometrics: an introduction*. Sade Publications, Inc.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4), 424.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal*, 6(1), 1-55.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99-114.

Huisman, M. (1999). *Item nonresponse: Occurrence, causes, and imputation of missing answers to test items*. Leiden: DSWO Press.

Kantrowitz, T., Kingry, D., Engelsted, J., Travinin, G., Lovering, E., & Gould, M. (in press). *Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: Validity and Subgroup Differences for Current and Alternative Composites*, AFRL-RH-WP-TR-2022-xxxx. Wright-Patterson Air Force Base, OH: Air Force Research Laboratory, 711 Human Performance Wing, Performance Optimization Branch.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3. Bask1). NY: Guilford.

Kline, T. J. (2005). *Classical test theory: Assumptions, equations, limitations, and item*

analyses. *Psychological testing: A practical approach to design and evaluation*, 91.

- Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y. C., & Dunleavy, E. (2007). *Literacy in Everyday Life: Results from the 2003 National Assessment of Adult Literacy*. NCES 2007-490. National Center for Education Statistics.
- Levene, H. (1961). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. . Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, (pp. 279-292). Stanford University Press.
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437-455.
- Morath, R., Parish, C., Lodato, M., Schwartz, K., & Weissmuller, J. (2008). *AFOQT form S effectiveness analyses: Subtest-level analyses and report*. AFCAPS-FR-2010-0011. Randolph AFB, TX: Air Force Personnel Center, Strategic Research and Assessments Branch.
- Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bifactor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bifactor models via Monte Carlo simulations. *Journal of Intelligence*, 3(1), 2-20.
- Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research In R. W. Robins R. C., Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 407–421).
- Nunnally, J. C. (1978). *Psychometric Theory*: 2d Ed. McGraw-Hill.
- Reise, S. P., & Waller, N. G. (2002). Item response theory for dichotomous assessment data. In F. Drasgow & N. Schmitt (Eds.), *Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data Analysis* (pp. 88-122). San Francisco: Jossey-Bass.
- Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not?. *International Journal of Testing*, 17(1), 74-104.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147.
- Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99-103. Chicago
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). Chapter 4: Cleaning up your act,

- Screening Data Prior to Analysis. In B. G. Tabachnik & L. S. Fidell (Eds.). *Using multivariate statistics* (Vol. 5, pp. 481-498). Pearson.
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273-1296.
- Thorndike, R. L. (1982). *Applied psychometrics*. Houghton Mifflin.
- US Equal Employment Opportunity Commission, US Civil Service Commission, US Department of Labor,. (1978). US Department of Justice.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.
- Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: an illustration with the MMPI. *Psychological Methods*, 5(1), 125.
- Walsh, J. L., Woolley, M., Brady, M., & Carretta, T. R. (in press). *Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: Item-Level Analysis*, AFRL-RH-WP-TR-2022-xxxx. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Performance Optimization Branch.
- Walsh, J. L., Woolley, M., Brady, M., & Carretta, T. R. (2022). *Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: The Situational Judgment Test*, AFRL-RH-WP-TR-2022-0080. Wright-Patterson AFB, OH: Air Force Research Laboratory, 711 Human Performance Wing, Performance Optimization Branch.
- Woolley, M. R., Walsh, J. L., Mann, K., Wilson, R., & Carretta, T. R. (in progress). *Self-Description Inventory-Officer (SDI-O): Item-, facet-, domain-, and composite-level analyses*, AFRL-RH-WP-TR-3022-xxxx. Air Force Research Laboratory, 711 Human Performance Wing, Performance Optimization Branch.

APPENDIX A: Stratified Samples Demographics

Sample Characteristic	OTS-CIV		OTS-AD		AECF		T1 USAFA		ROTC		ANG		AFRES	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Sex														
Male	5,622	70%	4,308	79%	132	78%	3,195	74%	9,148	72%	3,050	78%	1,111	75%
Female	2,397	30%	1,114	21%	38	22%	1,114	26%	3,604	28%	863	22%	367	25%
Unknown	7	0%	8	0%	0	0%	3	0%	13	0%	2	0%	3	0%
Race*														
American Indian/Alaska Native	374	5%	322	6%	8	5%	201	5%	728	6%	220	6%	86	6%
Asian	888	11%	472	9%	10	6%	494	11%	1,484	12%	259	7%	122	8%
Black/African American	1,143	14%	1,009	19%	36	21%	418	10%	1,662	13%	437	11%	300	20%
Native Hawaiian/Other Pacific Islander	165	2%	164	3%	7	4%	141	3%	332	3%	82	2%	27	2%
White	6,142	77%	3,983	73%	119	70%	3,641	84%	9,688	76%	3,243	83%	1,033	70%
Ethnicity														
Hispanic	1,144	14%	803	15%	33	19%	493	11%	2,093	16%	416	11%	230	16%
Non-Hispanic	6,761	84%	4,515	83%	134	79%	3,768	87%	10,388	81%	3,427	88%	1,218	82%
Unknown	121	2%	112	2%	3	2%	51	1%	284	2%	72	2%	33	2%
Sample Characteristic	OTS-CIV		OTS-AD		AECF		T2 USAFA		ROTC		ANG		AFRES	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Sex														
Male	5,160	71%	3,900	81%	129	77%	2,791	75%	8,834	71%	2,595	79%	877	72%
Female	2,115	29%	940	19%	38	23%	945	25%	3,552	29%	691	21%	337	28%
Unknown	9	0%	3	0%	1	1%	5	0%	27	0%	2	0%	1	0%
Race*														
American Indian/Alaska Native	329	5%	335	7%	10	6%	165	4%	753	6%	153	5%	67	6%
Asian	735	10%	395	8%	20	12%	419	11%	1,426	11%	223	7%	92	8%
Black/African American	1,052	14%	931	19%	38	23%	339	9%	1,606	13%	376	11%	235	19%
Native Hawaiian/Other Pacific Islander	157	2%	150	3%	3	2%	103	3%	312	3%	67	2%	33	3%
White	5,512	76%	3,473	72%	121	72%	3,160	84%	9,428	76%	2,703	82%	858	71%
Ethnicity														
Hispanic	1,044	14%	707	15%	22	13%	398	11%	2,060	17%	334	10%	180	15%
Non-Hispanic	6,083	84%	4,024	83%	140	83%	3,276	88%	10,086	81%	2,884	88%	1,012	83%
Unknown	157	2%	112	2%	6	4%	67	2%	267	2%	70	2%	23	2%

*The proportions for Race do not add to 100% because test-takers had an option to choose more than one race.

Note. N = Sample Size.

APPENDIX B. Statistics and Their Interpretation

Statistic	Interpretation
Confirmatory Factor Analysis (CFA)	CFA is a factor analytic technique that relies on <i>a priori</i> theory to examine a causal structure among variables.
Correlation (r)	A correlation is a relation between two (i.e., bivariate) or more (i.e., multivariate) variables. Positive correlations indicate that as one variable increases, the other variable increases in a predictable manner. Negative correlations indicate that as one variable increases, the other variable decreases in a predictable manner.
Cronbach's Alpha (α)	A measure of internal consistency (reliability) of a psychological scale.
CTT Item Difficulty (p -value)	In this context, a measure of difficulty. Represents the percentage of applicants who selected the correct answer. Higher p -values indicate easier items.
CTT Item Discriminability (Item-Total Correlation [ITC])	Correlation between a correct response to an item and the overall score on the assessment. Positive numbers indicate that people who do well on the item do well on the overall assessment. Negative numbers indicate that people who do poorly on the item do well on the assessment. In this sense, ITC is both a measure of internal consistency and discriminability.
Effect Size (Cohen's d)	Provides a size of the difference in the mean between the majority group and a specified minority group in standard deviation units (e.g., $d = .5$ indicates a half standard deviation difference in the means of two groups. In this report, d is expressed as a positive number and it is assumed that the majority group outperformed the minority group unless otherwise specified.
Flesch-Kincaid Reading Grade Level (RGL)	A readability formula used to assess approximate grade-level required to comprehend a selection of text. Lengthier sentences with longer words require more cognitive effort to comprehend and are assessed as higher grade level. The target reading level is 8 th grade to ensure at least 80% of Americans can comprehend the selected text.
Independent Samples t -test	A statistical procedure in which test-takers are assigned to one of two conditions.
Kurtosis	Kurtosis is a departure from normality in which the data is more peaked and densely distributed or flatter and less densely distributed than a normal distribution.
Linear Mixed Effects Model (LMEM)	LMEM is a model with one or more random effects. The common characteristic of these models is that they allow researchers to analyze data with one or more sources of non-independence. A variable is considered <i>fixed</i> when data have been gathered from all the levels of the variable that are of interest. A variable is considered <i>random</i> when it has many possible levels and when the researchers' interest is in all possible levels, but only a random sample of levels is included in the data.
Measurement Invariance (MI)	MI occurs when individuals with the same standing on the trait assessed by a test or a scale, but sampled from different groups, have equal expected observed scores.
Omission Rates	Omission rates refer to the percentage of test-takers who did not provide a response to a given item. Item omission may be indicative of difficulty, IER, or speededness of the assessment.
Paired Samples t -test	A statistical procedure used in a repeated measures design where each test-taker is measured at all measurement occasions.
Regression	A statistical technique which analyzes the statistical significance and relative importance of a predictor or collection of predictors in predicting a given outcome or outcomes.
R-Squared (R^2)	Indicates the proportion of variance in an outcome accounted for by a predictor or collection of predictors in a regression context.
Significance Testing	In this context, indicates the statistical significance of a given statistical test at a predetermined probability level. Indicates the probability that the results are due to chance characteristics alone rather than due to an actual effect. In this report, a traditional p -value of .05 was chosen for all statistical tests (i.e., $p < .05$) except where otherwise noted.
Skewness	Skewness is a departure from normality in which the data is clustered to one side of the distribution or the other.
Stability Analysis	Stability analysis is a way to examine whether or not the test scores change over time (which may indicate a test security breach).
Stepwise Regression	An empirical (rather than theoretical) regression procedure used to determine the most influential predictors of an outcome by adding and/or removing predictors based on their statistical significance level. Stepwise regression is particularly susceptible to overfitting of the data which may limit replicability. To limit overfitting, a more restrictive p -value of .01 was chosen for stepwise regression analyses.

APPENDIX C. Stratified Samples Descriptives

OTS-CIV										
Subtest	T1					T2				
	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
VA	8,025	15.16	4.57	-.26	-.51	7,284	15.04	4.68	-.27	-.54
AR	8,025	13.49	5.08	.10	-.68	7,284	13.30	5.27	.13	-.70
WK	8,023	15.48	5.70	-.19	-1.01	7,284	15.08	5.65	-.15	-.99
MK	8,023	14.00	5.26	.12	-.77	7,282	14.11	5.11	.13	-.70
RC	8,023	17.01	4.09	-.51	-.16	7,284	17.69	4.39	-.76	.20
PS	8,020	10.41	3.98	.17	-.74	7,280	10.83	3.88	.13	-.67
TR	8,019	27.88	6.45	-.25	-.14	7,278	27.41	6.63	-.29	.05
IC	8,018	14.50	6.82	-.26	-.99	7,279	14.58	6.56	-.21	-1.06
BC	8,019	16.52	6.13	-.19	-.54	7,278	17.09	5.48	-.14	-.36
AI	8,009	9.21	4.43	.56	-.48	7,267	8.62	4.01	.71	.01
OTS-AD										
Subtest	T1					T2				
	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
VA	5,430	14.54	4.23	-.07	-.51	4,841	14.33	4.33	-.08	-.55
AR	5,430	12.48	4.66	.22	-.54	4,840	12.03	4.79	.32	-.36
WK	5,430	15.14	5.50	-.06	-1.02	4,840	14.77	5.36	-.04	-.92
MK	5,430	11.67	4.53	.40	-.27	4,840	11.82	4.40	.40	-.16
RC	5,429	17.17	3.86	-.47	-.18	4,837	17.25	4.01	-.53	-.08
PS	5,425	8.80	3.46	.48	-.13	4,836	9.47	3.36	.33	-.22
TR	5,424	27.93	6.34	-.25	-.15	4,835	27.62	6.46	-.24	.11
IC	5,424	14.02	6.71	-.21	-1.02	4,832	14.01	6.40	-.12	-1.03
BC	5,425	15.79	6.00	-.08	-.56	4,836	16.72	5.42	-.09	-.34
AI	5,416	9.20	3.75	.46	-.22	4,827	8.71	3.56	.51	-.13
AECF										
Subtest	T1					T2				
	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
VA	170	13.96	4.47	-.06	-.60	168	13.66	4.34	.08	-.47
AR	170	11.52	4.78	.19	-.35	168	12.01	4.89	.19	-.67
WK	170	13.46	5.78	.18	-1.05	168	13.71	5.94	.23	-1.16
MK	170	12.00	4.49	.33	-.21	168	12.05	4.09	.08	-.71
RC	170	16.16	3.82	-.32	-.50	168	16.42	4.50	-.39	-.52
PS	170	8.58	3.49	.53	-.06	168	9.81	3.57	.23	-.36
TR	169	26.88	7.01	.06	-.65	168	26.45	6.75	-.19	-.57
IC	170	13.68	6.45	-.22	-.91	168	14.19	6.26	-.23	-.96
BC	170	14.89	6.45	.00	-.78	168	15.80	5.45	-.07	-.05
AI	169	7.91	3.48	.61	.10	166	7.87	3.40	.59	.16
USAFA										
Subtest	T1					T2				
	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
VA	4,312	15.84	3.81	-.29	-.23	3,741	16.03	3.96	-.35	-.25
AR	4,312	16.10	4.44	-.25	-.48	3,741	16.30	4.61	-.24	-.57
WK	4,310	16.20	5.04	-.30	-.77	3,737	15.78	4.94	-.25	-.78
MK	4,311	18.44	4.10	-.50	-.26	3,737	18.10	3.90	-.38	-.26
RC	4,311	18.14	3.69	-.80	.67	3,733	18.74	3.71	-.91	.98
PS	4,299	13.25	3.22	-.31	-.42	3,736	13.72	3.12	-.30	-.28
TR	4,299	28.43	5.54	-.09	.00	3,736	27.57	5.78	-.28	.62
IC	4,298	15.82	6.31	-.49	-.73	3,735	15.48	6.20	-.36	-.90
BC	4,299	13.72	5.22	.27	-.33	3,736	15.00	4.68	.06	-.12
AI	4,298	10.79	3.87	.26	-.64	3,734	10.20	3.49	.37	-.28

ROTC										
Subtest	T1					T2				
	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
VA	12,765	14.16	4.41	-.15	-.48	12,413	13.95	4.64	-.16	-.58
AR	12,764	12.90	5.02	.16	-.64	12,413	12.86	5.17	.16	-.69
WK	12,762	13.71	5.54	.09	-.96	12,413	13.49	5.37	.08	-.86
MK	12,763	14.72	5.30	-.05	-.84	12,413	14.59	5.04	-.01	-.76
RC	12,760	16.57	4.15	-.49	-.27	12,413	17.07	4.55	-.64	-.08
PS	12,759	10.91	3.93	.02	-.78	12,406	11.26	3.92	.00	-.69
TR	12,759	26.07	6.14	-.17	.10	12,405	25.56	6.36	-.22	.24
IC	12,755	14.39	6.83	-.26	-1.04	12,398	14.36	6.52	-.18	-1.09
BC	12,758	14.32	5.72	.08	-.47	12,405	15.27	5.08	-.02	-.27
AI	12,742	8.62	4.11	.66	-.22	12,391	8.31	3.75	.68	.03
ANG										
Subtest	T1					T2				
	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
VA	3,915	14.92	4.23	-.14	-.53	3,288	14.83	4.35	-.17	-.50
AR	3,915	12.89	4.78	.21	-.54	3,288	12.57	4.87	.20	-.55
WK	3,915	15.39	5.54	-.12	-1.02	3,288	15.19	5.38	-.14	-.90
MK	3,915	12.37	4.83	.33	-.42	3,288	12.53	4.53	.31	-.41
RC	3,915	17.49	3.83	-.55	.04	3,288	17.84	3.93	-.65	.11
PS	3,914	9.65	3.71	.36	-.38	3,288	10.27	3.59	.23	-.45
TR	3,914	28.52	6.39	-.19	-.28	3,288	28.18	6.64	-.39	.20
IC	3,912	15.16	6.83	-.38	-.96	3,288	15.12	6.60	-.27	-1.07
BC	3,914	16.34	5.99	-.15	-.55	3,288	17.07	5.32	-.17	-.37
AI	3,903	10.88	4.86	.27	-1.02	3,278	10.32	4.57	.35	-.83
AFRES										
Subtest	T1					T2				
	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>N</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
VA	1,481	14.57	4.35	-.15	-.50	1,215	14.32	4.42	-.16	-.52
AR	1,481	12.45	4.78	.22	-.51	1,215	12.33	5.11	.26	-.63
WK	1,481	15.13	5.78	-.13	-1.03	1,215	15.07	5.59	-.09	-1.07
MK	1,481	12.28	4.93	.32	-.56	1,215	12.57	4.80	.33	-.48
RC	1,481	17.15	4.00	-.51	-.12	1,215	17.39	4.05	-.54	-.06
PS	1,481	9.36	3.80	.32	-.55	1,215	9.97	3.73	.19	-.52
TR	1,480	27.87	6.49	-.12	-.46	1,215	27.49	6.55	-.23	-.08
IC	1,481	14.30	7.03	-.21	-1.08	1,215	14.64	6.81	-.23	-1.11
BC	1,481	15.56	6.41	-.03	-.68	1,215	16.61	5.61	-.12	-.44
AI	1,478	10.70	5.07	.30	-1.09	1,215	9.89	4.57	.41	-.73

Note. *N* = Sample Size; *M* = Mean; *SD* = Standard Deviation.

APPENDIX D. Stratified Samples Difficulty

Subtest	OTS-CIV <i>p</i> -values			OTS-AD <i>p</i> -values			AECF <i>p</i> -values			T1 USAFA <i>p</i> -values			ROTC <i>p</i> -values			ANG <i>p</i> -values			AFRES <i>p</i> -values		
	Min	Max	<i>M</i>	Min	Max	<i>M</i>	Min	Max	<i>M</i>	Min	Max	<i>M</i>	Min	Max	<i>M</i>	Min	Max	<i>M</i>	Min	Max	<i>M</i>
VA	.31	.80	.60	.27	.80	.58	.26	.78	.56	.29	.86	.63	.27	.78	.56	.30	.80	.59	.30	.81	.58
AR	.27	.78	.54	.19	.72	.49	.17	.65	.46	.28	.85	.64	.24	.69	.50	.22	.77	.51	.19	.76	.50
WK	.33	.85	.62	.31	.86	.60	.25	.81	.54	.38	.85	.64	.23	.82	.53	.34	.85	.61	.33	.86	.60
MK	.24	.79	.56	.13	.75	.46	.14	.73	.48	.38	.90	.73	.22	.78	.58	.14	.77	.49	.17	.77	.49
RC	.37	.94	.68	.35	.93	.68	.37	.92	.65	.45	.95	.72	.34	.93	.65	.37	.92	.70	.36	.92	.68
PS	.27	.85	.52	.24	.76	.43	.18	.80	.43	.31	.95	.66	.27	.87	.53	.26	.78	.48	.25	.76	.46
TR	.17	.97	.70	.19	.97	.70	.21	.96	.67	.17	.97	.71	.12	.96	.65	.21	.97	.71	.19	.97	.70
IC	.31	.77	.58	.29	.74	.56	.27	.72	.54	.35	.80	.63	.32	.72	.57	.36	.78	.61	.32	.76	.57
BC	.11	.88	.55	.08	.86	.53	.09	.86	.49	.05	.86	.46	.07	.84	.48	.09	.87	.54	.09	.85	.52
AI	.28	.85	.46	.21	.78	.46	.17	.76	.39	.30	.92	.54	.25	.79	.43	.37	.83	.54	.36	.82	.54

Subtest	OTS-CIV <i>p</i> -values			OTS-AD <i>p</i> -values			AECF <i>p</i> -values			T2 USAFA <i>p</i> -values			ROTC <i>p</i> -values			ANG <i>p</i> -values			AFRES <i>p</i> -values		
	Min	Max	<i>M</i>	Min	Max	<i>M</i>	Min	Max	<i>M</i>	Min	Max	<i>M</i>	Min	Max	<i>M</i>	Min	Max	<i>M</i>	Min	Max	<i>M</i>
VA	.33	.83	.60	.32	.82	.57	.27	.83	.55	.33	.91	.64	.32	.80	.55	.34	.83	.59	.32	.82	.57
AR	.27	.74	.53	.18	.70	.48	.22	.68	.48	.37	.87	.65	.27	.70	.50	.20	.73	.50	.20	.72	.49
WK	.31	.86	.60	.28	.84	.59	.26	.80	.54	.34	.88	.63	.21	.79	.53	.30	.86	.60	.30	.83	.60
MK	.24	.81	.56	.13	.74	.47	.05	.74	.48	.37	.95	.72	.21	.80	.57	.15	.76	.50	.18	.78	.50
RC	.37	.92	.71	.31	.93	.68	.31	.87	.66	.45	.93	.75	.35	.89	.67	.33	.93	.71	.34	.93	.69
PS	.27	.83	.54	.23	.76	.47	.19	.78	.49	.35	.96	.68	.27	.85	.55	.26	.78	.51	.26	.77	.49
TR	.16	.97	.68	.18	.96	.69	.12	.98	.66	.13	.97	.69	.12	.96	.64	.20	.97	.71	.18	.97	.69
IC	.32	.78	.58	.29	.76	.56	.30	.77	.57	.35	.80	.62	.33	.76	.57	.36	.79	.61	.35	.78	.59
BC	.20	.91	.57	.18	.92	.56	.12	.90	.52	.15	.91	.50	.16	.90	.51	.19	.92	.57	.19	.90	.55
AI	.21	.71	.43	.21	.78	.43	.15	.70	.39	.20	.94	.51	.23	.73	.41	.31	.78	.52	.28	.80	.50

Note. Min = Minimum; Max = Maximum; *M* = Mean. For subtest-specific *N*, please refer to the Attachment 2.

APPENDIX E. Stratified Samples Percentiles

OTS-CIV																										
Subtest	T1													T2												
	1	5	10	20	30	40	50	60	70	80	90	95	99	1	5	10	20	30	40	50	60	70	80	90	95	99
VA	5	7	9	11	13	14	15	17	18	19	21	22	24	4	7	9	11	13	14	15	17	18	19	21	22	24
AR	3	5	7	9	10	12	13	15	16	18	20	22	24	3	5	6	8	10	12	13	15	16	18	21	22	24
WK	4	6	8	10	12	14	16	18	20	21	23	24	25	4	6	7	10	12	13	15	17	19	21	23	24	25
MK	4	6	7	9	11	12	14	15	17	19	22	23	25	4	6	8	9	11	12	14	15	17	19	21	23	25
RC	6	10	11	13	15	16	18	19	20	21	22	23	24	6	9	11	14	16	17	18	19	20	22	23	24	25
PS	3	4	5	7	8	9	10	11	13	14	16	17	19	3	5	6	7	8	10	11	12	13	14	16	17	19
TR	12	17	19	23	25	26	28	30	31	33	37	39	40	10	16	19	22	24	26	28	29	31	33	36	38	40
IC	1	3	5	8	11	13	15	17	19	21	23	24	25	2	3	5	8	11	13	15	17	19	21	23	24	25
BC	3	6	8	11	13	15	17	18	20	22	24	26	29	4	8	10	12	14	16	17	19	20	22	24	26	28
AI	2	3	4	5	6	7	8	10	11	13	16	18	20	2	3	4	5	6	7	8	9	10	12	15	17	19
OTS-AD																										
Subtest	T1													T2												
	1	5	10	20	30	40	50	60	70	80	90	95	99	1	5	10	20	30	40	50	60	70	80	90	95	99
VA	5	8	9	11	12	13	15	16	17	18	20	21	23	5	7	9	11	12	13	14	16	17	18	20	21	23
AR	3	5	7	8	10	11	12	13	15	17	19	21	23	3	5	6	8	9	10	12	13	14	16	19	21	24
WK	4	6	8	10	12	13	15	17	19	21	23	24	25	4	6	8	10	11	13	15	16	18	20	22	23	25
MK	3	5	6	8	9	10	11	12	14	16	18	20	23	3	5	6	8	9	10	11	13	14	16	18	20	23
RC	7	10	12	14	15	16	18	19	20	21	22	23	24	7	10	12	14	15	17	18	19	20	21	22	23	24
PS	2	4	5	6	7	8	8	9	10	12	14	15	18	3	4	5	7	7	8	9	10	11	12	14	15	18
TR	12	17	20	23	25	27	28	30	31	34	36	38	40	12	17	19	22	24	26	28	29	31	33	37	39	40
IC	1	3	4	7	10	12	15	16	19	21	23	24	25	1	3	5	8	10	12	14	16	18	20	23	24	25
BC	3	6	8	11	13	14	16	18	19	21	24	25	28	4	8	10	12	14	15	17	18	20	21	24	26	28
AI	2	4	5	6	7	8	9	10	11	12	14	16	19	2	4	4	6	7	7	8	9	10	12	14	15	18

AECF																										
Subtest	T1													T2												
	1	5	10	20	30	40	50	60	70	80	90	95	99	1	5	10	20	30	40	50	60	70	80	90	95	99
VA	5	6	8	10	12	13	14	15	16	18	20	22	23	5	7	8	10	11	13	14	15	16	17	19	21	23
AR	2	4	6	7	9	10	11	13	14	15	17	19	22	3	5	6	7	9	11	12	13	15	16	18	20	23
WK	4	5	6	8	9	11	14	15	16	19	22	23	24	4	5	6	8	9	11	13	15	18	20	22	24	25
MK	4	5	6	8	10	11	11	13	14	16	18	20	23	4	5	7	8	9	11	12	13	14	16	18	19	20
RC	7	10	11	13	14	16	16	17	18	20	21	22	22	5	9	10	13	14	15	17	18	20	21	22	23	24
PS	3	4	4	6	6	7	8	9	10	11	13	15	17	3	4	6	7	8	9	10	10	12	13	14	16	17
TR	13	16	18	20	23	25	27	28	30	33	37	39	40	13	15	17	20	24	25	26	28	30	33	35	37	39
IC	1	2	4	8	10	13	14	16	18	20	22	24	25	2	3	5	8	11	13	14	16	19	20	22	24	25
BC	3	4	7	9	11	13	15	17	19	21	23	25	28	3	6	9	11	13	15	16	17	19	20	22	25	28
AI	2	3	4	5	6	7	8	8	9	10	13	15	17	2	3	4	5	6	7	8	8	9	10	12	14	17
USAFA																										
Subtest	T1													T2												
	1	5	10	20	30	40	50	60	70	80	90	95	99	1	5	10	20	30	40	50	60	70	80	90	95	99
VA	7	9	11	13	14	15	16	17	18	19	21	22	23	6	9	11	13	14	15	16	17	18	20	21	22	24
AR	6	9	10	12	14	15	16	18	19	20	22	23	24	6	8	10	12	14	15	16	18	19	21	22	24	25
WK	5	7	9	11	13	15	17	18	20	21	23	24	25	5	7	9	11	13	15	16	18	19	21	22	23	24
MK	8	11	13	15	16	18	19	20	21	22	24	24	25	9	11	13	15	16	17	18	19	21	22	23	24	25
RC	7	11	13	15	17	18	19	20	20	21	22	23	24	8	12	14	16	17	18	19	20	21	22	23	24	25
PS	6	8	9	10	12	13	13	14	15	16	17	18	19	6	8	10	11	12	13	14	15	16	17	18	18	20
TR	15	19	22	24	26	27	28	30	31	33	36	38	40	13	18	21	23	25	26	28	29	30	32	35	38	40
IC	1	4	6	10	13	15	17	19	20	22	24	24	25	2	4	7	10	12	14	16	18	20	22	23	24	25
BC	3	6	7	9	11	12	13	15	16	18	21	23	26	5	7	9	11	13	14	15	16	17	19	21	23	26
AI	4	5	6	7	8	9	10	12	13	14	16	18	19	3	5	6	7	8	9	10	11	12	13	15	17	19
ROTC																										
Subtest	T1													T2												
	1	5	10	20	30	40	50	60	70	80	90	95	99	1	5	10	20	30	40	50	60	70	80	90	95	99
VA	4	7	8	10	12	13	14	16	17	18	20	21	23	4	6	8	10	11	13	14	15	17	18	20	21	23
AR	3	5	6	8	10	11	13	14	16	18	20	22	24	3	5	6	8	10	11	13	14	16	18	20	22	24
WK	3	5	7	8	10	12	13	15	17	19	21	23	24	3	5	7	8	10	12	13	15	17	19	21	22	24
MK	4	6	8	10	11	13	15	16	18	20	22	23	25	4	6	8	10	12	13	14	16	18	19	21	23	25
RC	6	9	11	13	15	16	17	18	19	20	22	23	24	5	9	11	13	15	17	18	19	20	21	22	23	24
PS	3	5	6	7	8	10	11	12	13	15	16	17	19	3	5	6	8	9	10	11	12	14	15	17	18	19
TR	10	16	18	21	23	25	26	28	29	31	34	36	39	9	15	17	20	22	24	26	27	29	31	34	36	40
IC	1	3	4	7	10	13	15	17	19	21	23	24	25	2	4	5	8	10	13	15	17	19	21	23	24	25
BC	2	5	7	9	11	13	14	16	17	19	22	24	27	4	7	8	11	13	14	15	17	18	20	22	24	27
AI	2	3	4	5	6	7	8	9	10	12	15	17	19	2	3	4	5	6	7	8	9	10	11	14	16	18

ANG																										
Subtest	T1													T2												
	1	5	10	20	30	40	50	60	70	80	90	95	99	1	5	10	20	30	40	50	60	70	80	90	95	99
VA	5	8	9	11	13	14	15	16	17	19	20	22	23	5	8	9	11	13	14	15	16	17	19	21	22	23
AR	3	6	7	9	10	11	13	14	15	17	20	21	24	3	5	6	8	10	11	12	14	15	17	19	21	24
WK	4	6	8	10	12	14	15	17	19	21	23	24	25	4	6	8	10	12	14	15	17	19	21	22	24	25
MK	3	5	6	8	9	11	12	13	15	17	19	21	24	4	6	7	8	10	11	12	14	15	16	19	20	24
RC	8	11	12	14	16	17	18	19	20	21	22	23	24	7	10	12	15	16	17	18	19	20	21	22	23	25
PS	3	4	5	6	7	8	9	10	11	13	15	16	18	3	5	6	7	8	9	10	11	12	13	15	17	19
TR	13	18	20	23	25	27	28	30	32	34	38	39	40	10	17	20	23	25	27	28	30	32	34	37	39	40
IC	1	3	5	8	11	14	16	18	20	22	24	24	25	2	4	5	9	11	13	16	18	20	22	24	24	25
BC	3	6	8	11	13	15	17	18	20	22	24	26	28	5	8	10	13	14	16	17	19	20	22	24	26	28
AI	2	4	5	6	8	9	10	12	14	16	18	19	20	2	4	5	6	7	8	10	11	13	15	17	19	20
AFRES																										
Subtest	T1													T2												
	1	5	10	20	30	40	50	60	70	80	90	95	99	1	5	10	20	30	40	50	60	70	80	90	95	99
VA	5	7	9	11	12	13	15	16	17	19	20	21	23	5	7	8	10	12	13	15	16	17	18	20	21	23
AR	3	5	6	8	10	11	12	13	15	17	19	21	23	3	5	6	8	9	11	12	13	15	17	20	22	24
WK	4	6	7	9	12	13	15	17	19	21	23	24	25	4	6	8	10	11	13	15	17	19	21	23	23	25
MK	3	5	6	8	9	11	12	13	15	17	19	21	24	4	5	7	8	10	11	12	13	15	17	19	21	24
RC	7	10	12	14	15	16	18	19	20	21	22	23	24	7	10	12	14	16	17	18	19	20	21	22	23	24
PS	2	4	5	6	7	8	9	10	11	13	15	16	18	3	4	5	7	8	9	10	11	12	13	15	17	18
TR	13	17	19	22	24	26	28	30	31	34	37	39	40	12	17	19	22	24	26	28	29	31	33	36	39	40
IC	1	2	4	7	10	12	15	17	19	22	23	24	25	1	3	5	8	10	13	15	18	20	22	23	24	25
BC	2	5	7	10	12	14	16	17	19	21	24	26	28	4	7	9	12	14	15	17	18	20	21	24	26	28
AI	2	4	5	6	7	8	10	11	14	16	18	19	20	2	4	4	6	7	8	9	10	12	14	17	18	20

Note. For subtest-specific *N*, please refer to the Attachment 2.

APPENDIX F. Stratified Samples Speededness

OTS-CIV									
Subtest	T1 (% Item Omissions)								<i>M</i>
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	
VA	.47	.52	.76	2.53	6.83				2.22
AR	.61	1.37	1.64	3.28	7.36				2.85
WK	.46	.59	.77	1.54	3.52				1.38
MK	.85	.63	1.47	1.78	4.18				1.78
RC	.10	.08	.09	.20	.66				.23
PS	.13	.26	.16	.51					.26
TR	.10	.16	.36	1.68	6.64	16.25	27.73	36.38	11.16
IC	.25	.50	3.08	10.60	20.91				7.07
BC	.17	.13	1.15	4.86	13.92	24.02			7.38
AI	.28	.23	.39	1.55					.61
Subtest	T2 (% Item Omissions)								<i>M</i>
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	
VA	.43	.81	.95	2.15	6.41				2.15
AR	.96	1.20	1.67	3.72	7.33				2.98
WK	.82	.60	.88	1.57	3.33				1.44
MK	1.14	1.20	1.25	2.51	5.01				2.22
RC	.11	.10	.11	.25	.60				.23
PS	.10	.25	.20	.45					.25
TR	.14	.17	.48	1.88	7.04	17.19	28.49	36.37	11.47
IC	.24	.47	3.12	10.47	21.15				7.09
BC	.14	.11	.43	3.63	13.59	24.72			7.10
AI	.39	.36	.62	1.79					.79
OTS-AD									
Subtest	T1 (% Item Omissions)								<i>M</i>
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	
VA	.29	.27	.48	1.75	5.94				1.75
AR	.37	.87	1.04	2.55	5.58				2.08
WK	.31	.31	.39	1.00	2.27				.85
MK	.47	.33	.90	1.34	3.07				1.22
RC	.11	.07	.06	.13	.43				.16
PS	.08	.15	.11	.38					.18
TR	.10	.14	.29	1.38	5.50	13.25	23.66	31.44	9.47
IC	.18	.31	2.53	8.78	17.28				5.82
BC	.20	.13	.94	4.17	11.78	20.74			6.33
AI	.06	.11	.20	.74					.28
Subtest	T2 (% Item Omissions)								<i>M</i>
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	
VA	.25	.48	.58	1.71	5.26				1.66
AR	.45	.64	.95	3.00	6.11				2.23
WK	.38	.28	.43	.93	2.71				.95
MK	.54	.65	.65	1.98	4.05				1.57
RC	.07	.06	.03	.10	.31				.11
PS	.05	.06	.05	.25					.10
TR	.12	.15	.28	1.32	4.92	13.41	23.40	30.92	9.32
IC	.14	.46	2.50	8.58	17.38				5.81
BC	.14	.18	.51	3.13	11.16	20.99			6.02
AI	.16	.16	.30	1.02					.41

AECF									
Subtest	T1								
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	M
VA	.35	.00	.35	1.29	3.88				1.18
AR	.35	.82	.47	2.00	4.47				1.62
WK	.12	.24	.59	1.65	3.88				1.29
MK	1.06	.71	.59	.47	2.71				1.11
RC	.00	.00	.12	.24	.00				.07
PS	.12	.12	.12	.59					.24
TR	.24	.24	.00	1.53	7.29	16.47	24.59	30.59	10.12
IC	.12	.47	2.35	10.00	17.18				6.02
BC	.47	.00	2.82	7.41	15.41	21.65			7.96
AI	.12	.24	.59	1.54					.62
Subtest	T2 (% Item Omissions)								
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	M
VA	1.07	.95	.95	2.38	8.10				2.69
AR	2.02	1.07	1.19	2.86	9.17				3.26
WK	1.19	.24	.71	.95	3.92				1.40
MK	1.19	2.26	.83	1.90	7.62				2.76
RC	.11	.36	.00	.24	.59				.26
PS	.12	.12	.12	.71					.27
TR	.12	.12	.48	2.86	7.26	15.71	29.17	38.81	11.82
IC	.24	.95	2.62	8.33	21.55				6.74
BC	.12	.12	.12	4.88	15.36	27.26			7.98
AI	.36	.48	.72	.96					.63
USAFA									
Subtest	T1 (% Item Omissions)								
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	M
VA	.19	.32	.45	2.20	8.30				2.29
AR	.17	.47	.60	1.57	4.53				1.47
WK	.16	.15	.19	.49	.93				.38
MK	.22	.21	.32	.52	1.33				.52
RC	.06	.11	.08	.07	.21				.11
PS	.03	.07	.11	.07					.07
TR	.09	.16	.15	.66	3.93	13.89	28.06	38.60	10.69
IC	.24	.21	1.40	6.56	15.63				4.81
BC	.38	.13	.60	3.62	12.99	25.38			7.19
AI	.07	.10	.08	.19					.11
Subtest	T2 (% Item Omissions)								
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	M
VA	.12	.36	.49	1.49	6.11				1.71
AR	.28	.37	.51	1.48	4.06				1.34
WK	.35	.13	.28	.48	1.17				.48
MK	.33	.34	.21	.66	1.51				.61
RC	.12	.11	.09	.10	.10				.10
PS	.03	.03	.07	.10					.06
TR	.08	.17	.25	.87	4.52	14.95	28.78	37.55	10.90
IC	.19	.20	1.65	8.03	18.11				5.64
BC	.21	.08	.17	2.72	13.13	25.11			6.90
AI	.13	.13	.13	.31					.18

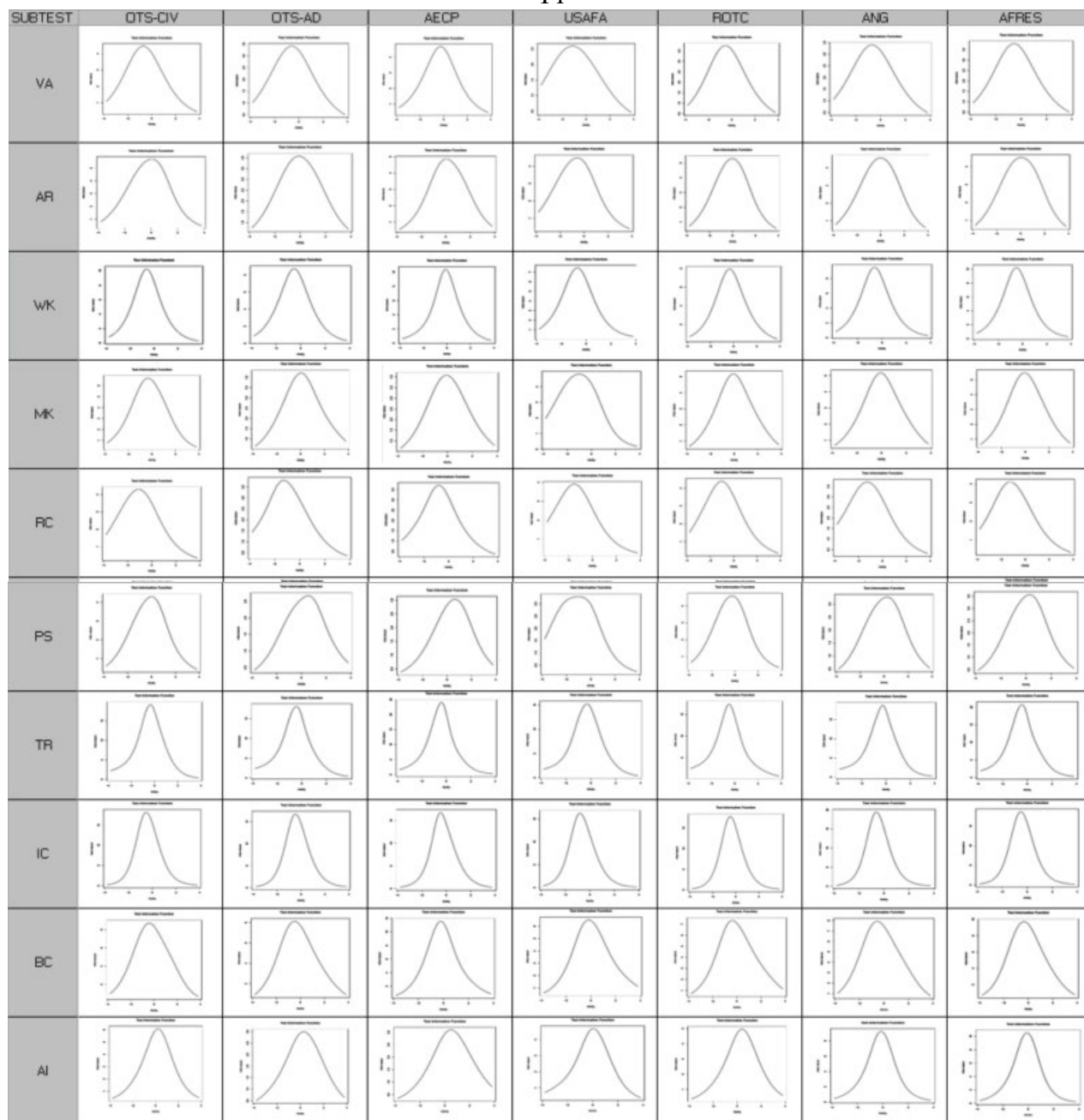
ROTC									
Subtest	T1 (% Item Omissions)								
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	M
VA	.42	.53	.73	2.04	5.27				1.80
AR	.60	1.07	1.35	2.43	5.14				2.12
WK	.55	.63	.77	1.60	2.76				1.26
MK	.61	.49	1.03	1.32	2.80				1.25
RC	.13	.11	.08	.17	.43				.18
PS	.09	.15	.15	.24					.16
TR	.17	.19	.47	2.05	6.77	15.92	25.37	32.85	10.47
IC	.19	.36	1.98	6.80	14.42				4.75
BC	.22	.15	.92	4.16	11.69	20.54			6.28
AI	.15	.19	.26	.91					.38
Subtest	T2 (% Item Omissions)								
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	M
VA	.32	.63	.88	1.67	4.73				1.64
AR	.72	.87	1.25	2.69	5.34				2.17
WK	.81	.66	.97	1.56	2.86				1.37
MK	.70	.78	.86	1.73	3.67				1.55
RC	.13	.14	.13	.23	.46				.22
PS	.09	.16	.12	.28					.16
TR	.14	.20	.58	2.43	7.50	16.51	26.19	33.03	10.82
IC	.22	.42	2.15	7.22	15.09				5.02
BC	.20	.12	.50	3.46	11.25	20.54			6.01
AI	.18	.24	.44	1.13					.50
ANG									
Subtest	T1 (% Item Omissions)								
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	M
VA	.31	.40	.48	1.85	5.27				1.66
AR	.44	.80	.97	2.17	5.35				1.95
WK	.28	.27	.35	.86	1.93				.74
MK	.52	.33	.90	1.13	2.72				1.12
RC	.08	.10	.08	.12	.46				.17
PS	.09	.17	.07	.36					.17
TR	.09	.17	.23	1.05	4.33	11.53	19.92	27.07	8.05
IC	.25	.38	1.92	6.72	13.81				4.61
BC	.18	.12	.79	3.12	9.18	17.36			5.13
AI	.13	.06	.19	.86					.31
Subtest	T2 (% Item Omissions)								
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	M
VA	.32	.58	.65	1.75	5.23				1.71
AR	.60	.88	1.31	2.70	5.62				2.22
WK	.43	.38	.50	.74	2.40				.89
MK	.55	.70	.79	1.56	3.47				1.41
RC	.13	.14	.13	.23	.46				.22
PS	.10	.09	.10	.28					.14
TR	.12	.12	.48	2.86	7.26	15.71	29.17	38.81	11.82
IC	.09	.33	1.82	7.06	14.56				4.77
BC	.16	.06	.33	2.35	9.26	18.28			5.08
AI	.14	.12	.30	.95					.38

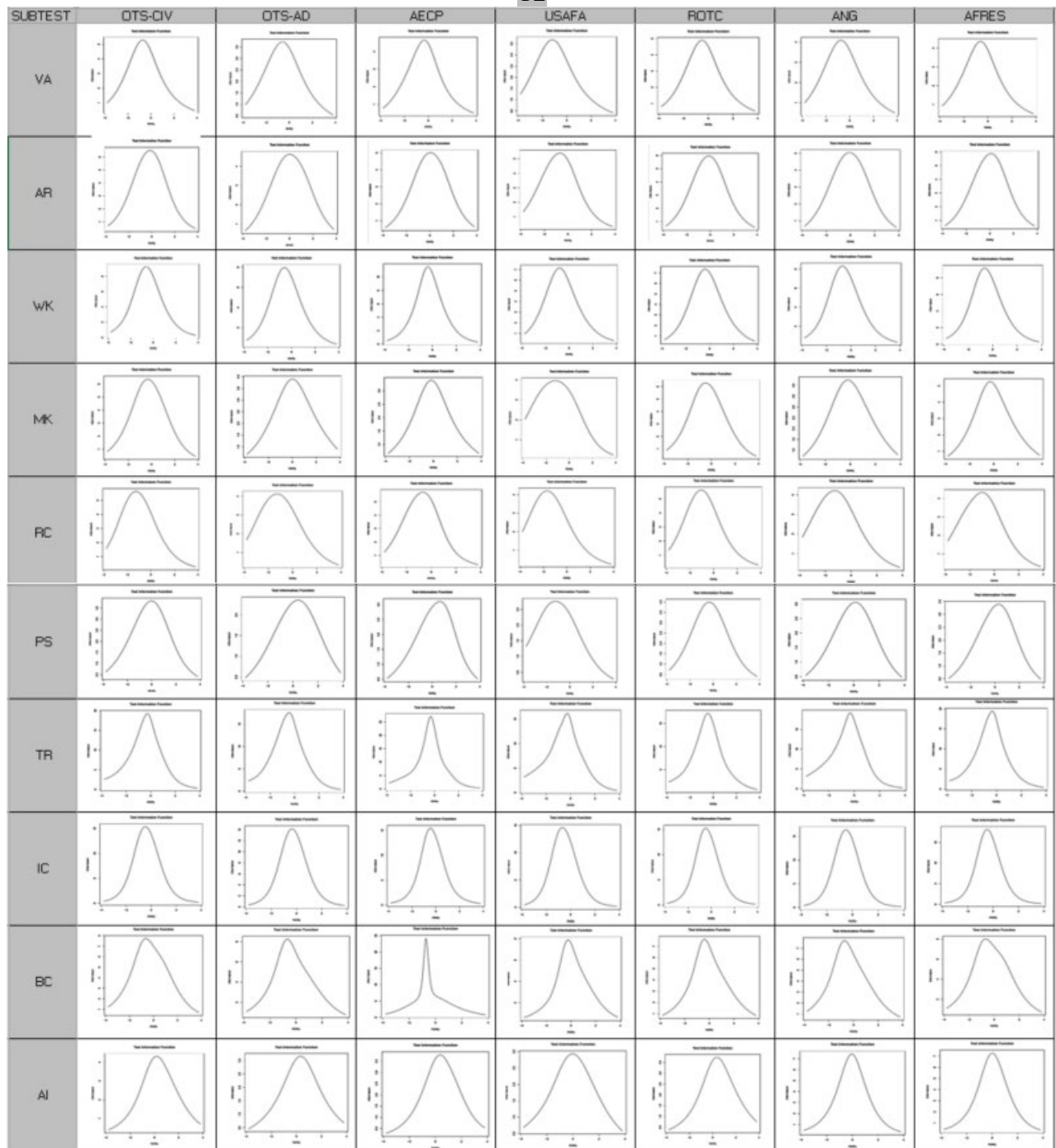
AFRES									
Subtest	T1 (% Item Omissions)								
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	<i>M</i>
VA	.38	.57	.90	2.38	6.78				2.20
AR	.46	1.28	1.63	3.34	6.48				2.64
WK	.38	.45	.59	1.27	2.67				1.07
MK	.62	.58	1.30	1.96	3.66				1.62
RC	.08	.11	.08	.08	.50				.17
PS	.16	.19	.20	.65					.30
TR	.12	.08	.39	1.78	5.91	14.64	25.04	31.74	9.96
IC	.28	.62	2.93	9.89	18.58				6.46
BC	.20	.11	1.35	5.85	13.48	22.17			7.19
AI	.12	.07	.23	1.10					.38
Subtest	T2 (% Item Omissions)								
	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25	26 to 30	31 to 35	36 to 40	<i>M</i>
VA	.40	.76	1.00	2.44	6.09				2.14
AR	.82	1.10	1.35	3.64	6.21				2.62
WK	.82	.51	.66	.87	2.47				1.07
MK	.97	1.05	.97	2.12	5.20				2.06
RC	.16	.08	.08	.13	.58				.21
PS	.03	.05	.13	.49					.18
TR	.08	.07	.41	1.66	6.45	14.44	23.97	32.00	9.88
IC	.20	.36	2.35	8.31	16.58				5.56
BC	.07	.07	.71	3.59	12.02	21.60			6.34
AI	.15	.23	.49	1.37					.56

Note. *M* = Mean. For subtest-specific *N*, please refer to the Attachment 2.

APPENDIX G. Stratified Samples Test Information Functions

T1





APPENDIX H. Stratified Samples Discriminability

Subtest	T1																				
	OTS-CIV ITC			OTS-AD ITC			AECF ITC			USAFA ITC			ROTC ITC			ANG ITC			AFRES ITC		
	Min	Max	M	Min	Max	M	Min	Max	M	Min	Max	M	Min	Max	M	Min	Max	M	Min	Max	M
VA	.27	.49	.39	.20	.48	.36	.17	.58	.37	.22	.42	.33	.23	.50	.37	.25	.46	.36	.22	.49	.37
AR	.28	.57	.43	.24	.48	.39	.17	.53	.40	.27	.54	.39	.26	.54	.42	.26	.49	.40	.25	.51	.40
WK	.35	.63	.48	.34	.60	.47	.21	.68	.48	.27	.60	.43	.31	.61	.46	.35	.60	.47	.37	.62	.49
MK	.29	.55	.44	.25	.53	.38	.22	.51	.38	.27	.55	.39	.29	.54	.45	.28	.56	.41	.23	.56	.41
RC	.22	.52	.37	.21	.49	.35	.18	.53	.34	.23	.48	.35	.24	.51	.38	.22	.49	.35	.21	.48	.36
PS	.25	.60	.42	.20	.52	.37	.18	.50	.37	.24	.58	.38	.23	.59	.42	.22	.54	.39	.25	.55	.39
TR	.15	.69	.42	.18	.68	.41	.11	.70	.42	.16	.66	.38	.19	.65	.40	.14	.71	.41	.14	.71	.41
IC	.36	.66	.58	.35	.65	.57	.33	.68	.54	.26	.65	.55	.30	.67	.57	.33	.68	.59	.37	.68	.59
BC	.29	.57	.45	.29	.57	.44	.22	.65	.47	.22	.53	.39	.28	.56	.42	.32	.54	.44	.34	.59	.47
AI	.30	.61	.47	.29	.54	.40	.22	.52	.38	.21	.60	.41	.27	.60	.44	.32	.67	.51	.33	.70	.53
Subtest	T2																				
	OTS-CIV ITC			OTS-AD ITC			AECF ITC			USAFA ITC			ROTC ITC			ANG ITC			AFRES ITC		
	Min	Max	M	Min	Max	M	Min	Max	M	Min	Max	M	Min	Max	M	Min	Max	M	Min	Max	M
VA	.30	.55	.40	.27	.51	.36	.14	.55	.36	.23	.47	.35	.25	.55	.38	.24	.50	.37	.27	.54	.37
AR	.26	.58	.44	.24	.50	.40	.18	.53	.41	.29	.57	.40	.27	.56	.43	.24	.53	.41	.24	.59	.42
WK	.30	.61	.48	.27	.60	.45	.20	.67	.49	.25	.57	.43	.24	.60	.45	.28	.59	.45	.30	.59	.47
MK	.30	.59	.43	.28	.53	.37	.17	.52	.35	.26	.52	.37	.28	.61	.43	.25	.52	.38	.30	.56	.40
RC	.26	.54	.41	.23	.49	.37	.15	.59	.40	.24	.48	.37	.24	.53	.41	.19	.50	.37	.19	.51	.38
PS	.25	.55	.41	.19	.47	.35	.25	.51	.37	.17	.54	.37	.25	.53	.41	.18	.52	.38	.19	.53	.39
TR	.20	.69	.42	.20	.67	.41	.04	.70	.41	.20	.66	.39	.19	.65	.41	.19	.68	.42	.18	.68	.41
IC	.36	.65	.56	.33	.65	.54	.29	.67	.53	.27	.65	.53	.30	.65	.55	.35	.68	.56	.39	.68	.57
BC	.23	.59	.41	.20	.60	.41	.18	.65	.41	.17	.60	.36	.21	.61	.39	.22	.58	.40	.21	.58	.42
AI	.32	.56	.42	.27	.54	.38	.20	.58	.38	.21	.50	.37	.26	.57	.40	.31	.66	.48	.29	.65	.48

Note. Min = Minimum; Max = Maximum; M = Mean. For subtest-specific N, please refer to the Attachment 2.

APPENDIX I. Stratified Samples Correlations

OTS-CIV																					
Subtest	T1										T2										
	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	
VA																					
AR	.52										.53										
WK	.72	.43									.72	.43									
MK	.46	.72	.38								.49	.75	.39								
RC	.63	.47	.67	.41							.67	.50	.66	.46							
PS	.52	.53	.50	.60	.50						.51	.54	.49	.59	.48						
TR	.31	.43	.21	.39	.26	.26					.32	.43	.20	.41	.30	.25					
IC	.43	.46	.33	.42	.41	.49	.39				.42	.45	.37	.43	.42	.47	.37				
BC	.42	.51	.30	.47	.34	.41	.51	.56			.41	.52	.31	.47	.37	.39	.50	.55			
AI	.36	.34	.34	.31	.40	.46	.27	.59	.38		.32	.31	.32	.29	.35	.43	.26	.59	.34		
OTS-AD																					
Subtest	T1										T2										
	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	
VA																					
AR	.48										.44										
WK	.66	.37									.66	.35									
MK	.41	.63	.30								.41	.67	.30								
RC	.57	.41	.63	.31							.58	.41	.60	.34							
PS	.48	.43	.47	.47	.41						.46	.44	.45	.47	.39						
TR	.29	.43	.18	.37	.24	.19					.26	.38	.16	.34	.25	.20					
IC	.39	.40	.31	.34	.39	.42	.39				.39	.41	.34	.37	.36	.43	.37				
BC	.36	.48	.24	.41	.29	.34	.51	.54			.35	.47	.26	.41	.30	.33	.49	.53			
AI	.36	.30	.36	.25	.40	.43	.26	.54	.35		.35	.28	.37	.24	.34	.44	.25	.54	.34		

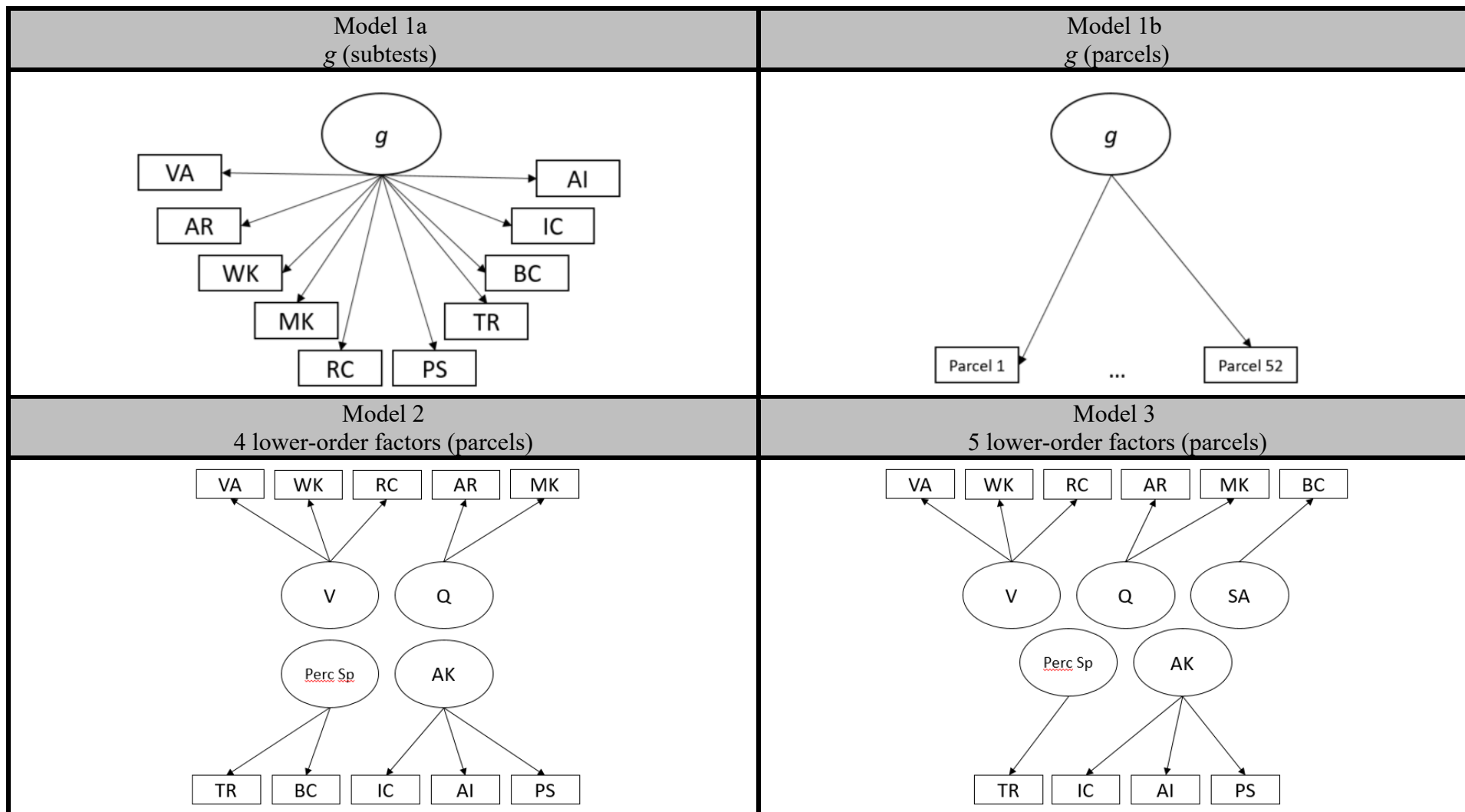
AECF																					
Subtest	T1										T2										
	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	
VA																					
AR	.62										.55										
WK	.70	.50									.72	.46									
MK	.48	.66	.37								.47	.64	.35								
RC	.63	.48	.65	.36							.71	.53	.76	.41							
PS	.42	.46	.46	.51	.39						.48	.56	.48	.61	.51						
TR	.29**	.34	.13 [†]	.28**	.16 [†]	.21*					.28**	.40	.14 [†]	.38	.29**	.19*					
IC	.41	.48	.33	.46	.29**	.48	.41				.41	.42	.40	.46	.45	.49	.34				
BC	.42	.49	.24	.44	.27**	.33	.48	.69			.39	.53	.31	.42	.33	.33	.36	.55			
AI	.35	.36	.43	.27	.35	.40	.19*	.52	.34		.25**	.26**	.32	.30**	.32	.43	.23**	.51	.37		
USAFA																					
Subtest	T1										T2										
	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	
VA																					
AR	.41										.43										
WK	.60	.36									.63	.36									
MK	.40	.65	.35								.40	.69	.36								
RC	.46	.37	.55	.36							.53	.41	.56	.39							
PS	.45	.46	.48	.47	.44						.42	.46	.46	.45	.41						
TR	.17	.31	.09	.30	.14	.14					.22	.35	.12	.33	.19	.14					
IC	.28	.34	.25	.31	.30	.43	.22				.28	.36	.28	.33	.25	.41	.24				
BC	.29	.38	.25	.36	.24	.33	.30	.41			.30	.41	.26	.37	.23	.32	.30	.44			
AI	.35	.31	.39	.27	.41	.48	.13	.52	.29		.32	.27	.35	.27	.30	.43	.16	.52	.28		
ROTC																					
Subtest	T1										T2										
	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	
VA																					
AR	.53										.55										
WK	.69	.46									.71	.45									
MK	.51	.73	.44								.53	.75	.43								
RC	.61	.49	.65	.46							.67	.52	.66	.50							
PS	.55	.57	.53	.62	.52						.54	.56	.51	.61	.52						
TR	.30	.43	.21	.42	.26	.27					.29	.41	.19	.42	.29	.26					
IC	.39	.44	.32	.43	.39	.50	.33				.39	.44	.35	.44	.40	.49	.33				
BC	.39	.48	.29	.48	.34	.42	.47	.53			.38	.48	.31	.48	.36	.40	.45	.50			
AI	.32	.32	.31	.30	.37	.45	.22	.57	.35		.32	.32	.32	.31	.35	.44	.21	.57	.33		

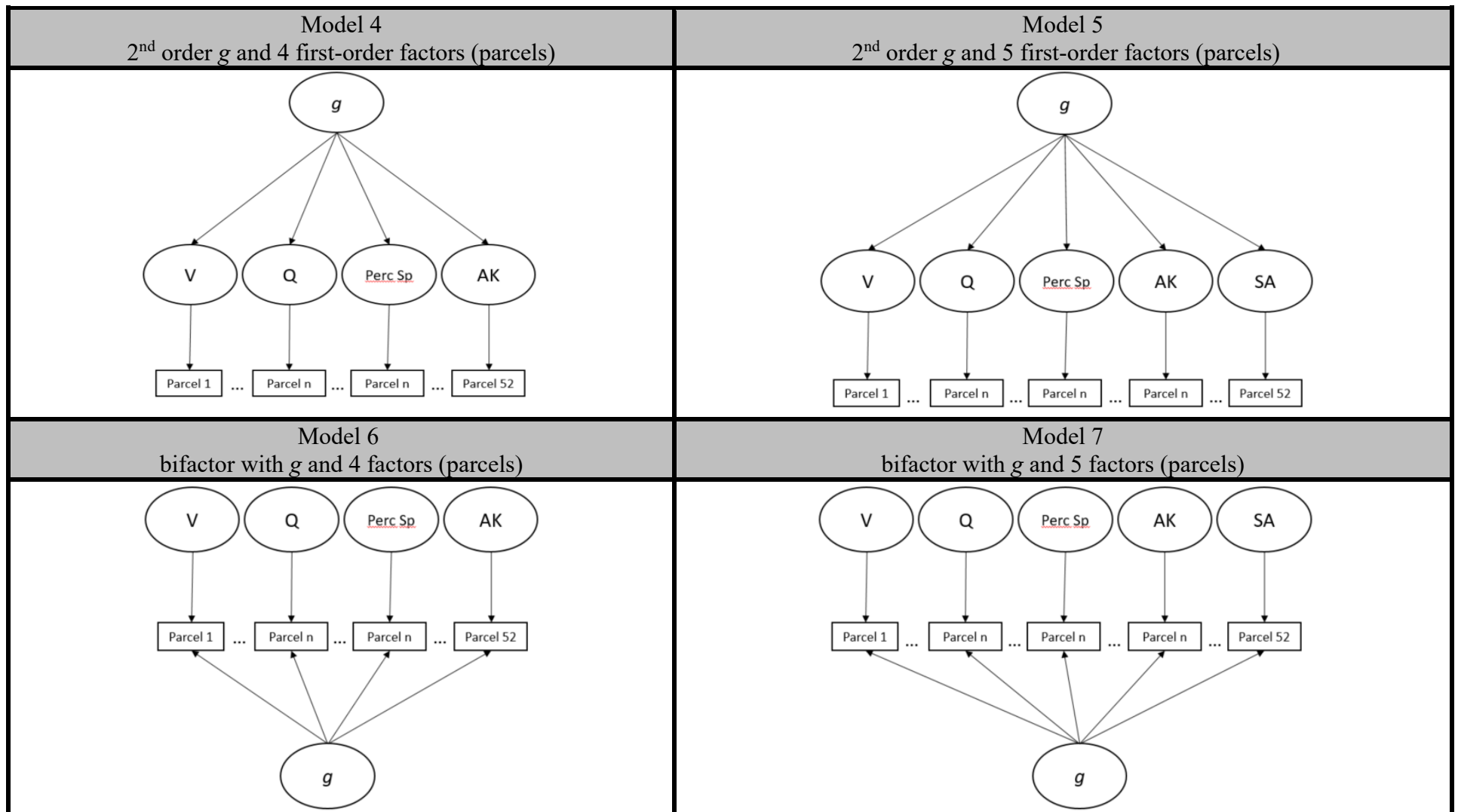
ANG																					
Subtest	T1										T2										
	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	
VA																					
AR	.50										.48										
WK	.66	.40									.65	.34									
MK	.43	.66	.32								.43	.71	.31								
RC	.57	.43	.62	.35							.60	.44	.59	.39							
PS	.45	.46	.44	.54	.42						.48	.49	.45	.53	.44						
TR	.26	.42	.11	.34	.19	.18					.27	.41	.15	.36	.25	.23					
IC	.36	.42	.26	.39	.34	.45	.35				.37	.43	.28	.40	.35	.45	.36				
BC	.37	.47	.23	.43	.27	.35	.50	.53			.34	.48	.22	.43	.29	.36	.47	.52			
AI	.30	.30	.26	.31	.34	.43	.23	.63	.37												
AFRES																					
Subtest	T1										T2										
	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	VA	AR	WK	MK	RC	PS	TR	IC	BC	AI	
VA																					
AR	.50										.51										
WK	.69	.40									.70	.39									
MK	.45	.68	.35								.44	.72	.31								
RC	.60	.46	.64	.40							.61	.48	.63	.39							
PS	.49	.48	.49	.56	.46						.43	.48	.45	.53	.43						
TR	.34	.44	.23	.37	.29	.27					.29	.40	.20	.34	.27	.22					
IC	.40	.42	.32	.39	.41	.49	.37				.38	.46	.33	.41	.37	.47	.38				
BC	.43	.51	.29	.45	.37	.41	.52	.57			.37	.53	.30	.45	.34	.38	.46	.57			
AI	.35	.33	.36	.34	.41	.50	.28	.66	.43												

For AECF: † $p \leq .100$; * $p \leq .050$; ** $p \leq .010$. For the rest of the accession sources: All unmarked correlations are significant at $p \leq .010$.

Note. For subtest-specific N , please refer to the Attachment 2.

APPENDIX J. CFA Path Diagrams





APPENDIX K. Stratified Samples Subgroup Differences

OTS-CIV												
Subgroup	VA						PS					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	14.43	4.75		14.51	4.81		8.38	3.32		8.95	3.39	
Male	15.47	4.45	.23**	15.25	4.61	.16**	11.27	3.93	.77**	11.59	3.81	.72**
Black	12.28	4.55	.88**	11.96	4.55	.90**	7.87	3.41	.80**	8.55	3.44	.75**
White	16.01	4.22		15.89	4.32		10.95	3.88		11.35	3.79	
Asian	14.32	4.91		14.35	4.72		10.84	3.93		11.16	3.99	
White	15.95	4.24	.38**	15.82	4.34	.34**	10.9	3.90	.02	11.3	3.78	.04
White Hispanic	14.49	4.39		14.51	4.69		9.93	3.83		10.65	3.67	
White Non-Hispanic	16.09	4.21	.38**	15.98	4.30	.34**	11.01	3.89	.28**	11.36	3.80	.19**
Black Hispanic	13.12	4.60	.18†	12.36	4.77	.04	8.44	3.95	.13	8.96	3.37	.11
Black Non-Hispanic	12.31	4.52		12.17	4.65		7.97	3.45		8.59	3.44	
Subgroup	AR						TR					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>D</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	11.47	4.66		11.33	4.87		27.23	6.35		26.78	6.47	
Male	14.35	5.00	.59**	14.11	5.22	.54**	28.16	6.48	.14**	27.67	6.67	.14**
Black	10.08	4.38		9.71	4.44		23.93	6.23		23.2	6.45	
White	14.22	4.89	.86**	14.09	5.08	.87**	28.83	6.15	.80**	28.34	6.26	.82**
Asian	14.6	5.45		14.52	5.62		28.05	6.28		27.77	6.62	
White	14.12	4.90	.10*	14.01	5.09	.10*	28.75	6.19	.11*	28.25	6.29	.07
White Hispanic	12.66	4.76		12.7	4.78		27.12	6.23		26.88	6.22	
White Non-Hispanic	14.27	4.89	.33**	14.1	5.15	.27**	28.9	6.16	.29**	28.41	6.30	.24**
Black Hispanic	10.2	4.55		10	4.46		23.35	6.53		23.55	6.53	
Black Non-Hispanic	10.11	4.28	.02	9.89	4.43	.03	24.17	6.30	.13	23.31	6.50	.04
Subgroup	WK						IC					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	14.7	5.69		13.93	5.65		10.05	5.82		10.39	5.61	
Male	15.82	5.67	.20**	15.55	5.58	.29**	16.39	6.31	1.03**	16.3	6.14	.99**
Black	12.33	5.39		11.63	5.22		8.58	5.73		9.03	5.61	
White	16.42	5.45	.75**	16	5.42	.81**	15.81	6.41	1.14**	15.89	6.17	1.12**
Asian	14.15	6.05		13.91	5.66		13.62	6.62		13.94	6.15	
White	16.36	5.46	.40**	15.94	5.43	.37**	15.74	6.45	.33**	15.75	6.22	.29**
White Hispanic	14.84	5.43		14.67	5.40		13.56	6.59		14.22	6.42	
White Non-Hispanic	16.5	5.44	.30**	16.1	5.42	.26**	15.98	6.37	.38**	15.94	6.19	.28**
Black Hispanic	13.53	5.78		12.5	5.31		9.23	6.21		10.24	5.76	
Black Non-Hispanic	12.44	5.41	.20*	11.83	5.29	.13	9.24	6.08	.00	9.27	5.72	.17†

Subgroup	MK						BC					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	12.21	4.84	.50**	12.35	4.8	.50**	14.61	6.06	.45**	15.48	5.48	.42**
Male	14.76	5.24		14.83	5.06		17.34	5.98		17.75	5.34	
Black	11.09	4.53	.68**	11.09	4.29	.74**	11.35	5.43	1.11**	12.5	4.88	1.11**
White	14.49	5.08		14.7	4.95		17.64	5.7		18.12	5.11	
Asian	16.54	5.66	.42**	16.31	5.4	.34**	16.46	6.25	.19**	17.41	5.15	.11*
White	14.38	5.09		14.6	4.97		17.55	5.74		17.97	5.16	
White Hispanic	13.16	4.85	.27**	13.6	4.98	.22**	15.99	5.81	.30**	16.79	5.04	.26**
White Non-Hispanic	14.54	5.12		14.69	4.97		17.71	5.72		18.12	5.17	
Black Hispanic	11.13	4.84	.01	10.87	4.62	.08	12.2	6.11	.08	13.34	4.63	.19†
Black Non-Hispanic	11.09	4.6		11.23	4.35		11.72	5.72		12.42	4.97	
Subgroup	RC						AI					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	15.81	4.01	.43**	16.78	4.59	.30**	7.06	3.51	.73**	6.58	2.98	.76**
Male	17.52	4.01		18.07	4.25		10.12	4.47		9.46	4.08	
Black	14.18	4.07	.95**	14.47	4.67	1.05**	6.25	3.24	.86**	6.1	3.1	.83**
White	17.77	3.75		18.62	3.85		9.94	4.44		9.35	4.03	
Asian	15.93	4.32	.47**	16.08	4.8	.63**	8.17	3.77	.39**	7.4	3.57	.48**
White	17.72	3.78		18.57	3.88		9.9	4.44		9.31	4.03	
White Hispanic	16.65	3.99	.31**	17.31	4.32	.36**	8.39	4.1	.38**	8.02	3.54	.35**
White Non-Hispanic	17.83	3.73		18.71	3.83		10.08	4.46		9.43	4.05	
Black Hispanic	14.9	3.85	.12	15.15	5.01	.09	6.6	3.28	.04	6.1	3.03	.05
Black Non-Hispanic	14.42	4.18		14.72	4.69		6.46	3.43		6.25	3.09	

† $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M = Mean; SD = Standard Deviation; d = Cohen's d . Underlined are Cohen's d for which minority test-takers outperformed majority test-takers.

Total T1: $N_{\text{Female}} = 2,397$, $N_{\text{Male}} = 5,621$;

Total T2: $N_{\text{Female}} = 2,115$, $N_{\text{Male}} = 5,160$;

Total T1: $N_{\text{Black}} = 754$, $N_{\text{White}} = 5,456$;

Total T2: $N_{\text{Black}} = 700$, $N_{\text{White}} = 4,948$;

Total T1: $N_{\text{Asian}} = 521$, $N_{\text{White}} = 5,373$;

Total T2: $N_{\text{Asian}} = 437$, $N_{\text{White}} = 4,900$;

Total T1: $N_{\text{White Non-Hispanic}} = 5,307$, $N_{\text{White Hispanic}} = 766$;

Total T2: $N_{\text{White Non-Hispanic}} = 4,753$, $N_{\text{White Hispanic}} = 684$;

Total T1: $N_{\text{Black Non-Hispanic}} = 994$, $N_{\text{Black Hispanic}} = 124$;

Total T2: $N_{\text{Black Non-Hispanic}} = 911$, $N_{\text{Black Hispanic}} = 109$.

For subtest-specific N , please refer to the Attachment 2.

OTS-AD												
Subgroup	VA						PS					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	13.68	4.27	.26**	13.64	4.35	.20**	7.04	2.96	.66**	7.73	2.93	.66**
Male	14.76	4.20		14.49	4.31		9.25	3.44		9.89	3.32	
Black	12.28	3.85	.77**	12.51	4.09	.61**	6.93	2.69	.72**	7.81	2.91	.64**
White	15.38	4.06		15.07	4.20		9.36	3.48		9.93	3.36	
Asian	13.2	4.46	.53**	13.42	4.3	.39**	8.58	3.64	.21**	9.49	3.52	.12†
White	15.36	4.07		15.04	4.18		9.32	3.48		9.9	3.36	
White Hispanic	14.51	4.27	.22**	13.74	4.28	.34**	9.11	3.32	.06	9.42	3.08	.15**
White Non-Hispanic	15.4	4.03		15.16	4.16		9.31	3.49		9.93	3.38	
Black Hispanic	12.32	4.28	.07	13.2	4.100		7.43	2.92	.13	8.49	2.43	.19†
Black Non-Hispanic	12.6	3.82		12.59	4.12	.15	7.07	2.82		7.93	2.99	
Subgroup	AR						TR					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	10.56	4.42	.53**	9.95	4.3	.55**	26.69	6.59	.25**	27.01	6.40	.12**
Male	12.98	4.59		12.54	4.76		28.25	6.24		27.77	6.47	
Black	9.45	3.66	.89**	9.55	4.00	.71**	23.81	5.84	.90**	23.7	6.38	.82**
White	13.38	4.54		12.81	4.71		29.16	5.96		28.8	6.14	
Asian	12.68	4.97	.14*	13.06	5.26	.06	28.04	6.5	.17**	27.4	5.81	.22**
White	13.3	4.53		12.77	4.7		29.05	5.99		28.73	6.16	
White Hispanic	12.24	4.43	.25**	11.49	4.38	.29**	27.32	6.14	.32**	27.25	6.03	.26**
White Non-Hispanic	13.38	4.53		12.86	4.73		29.25	5.91		28.84	6.18	
Black Hispanic	9.9	3.89	.05	10.22	4.39	.15	24.61	6.37	.10	23.79	6.64	.06
Black Non-Hispanic	9.69	3.81		9.61	4.13		24.02	6.00		24.17	6.34	
Subgroup	WK						IC					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	14.14	5.54	.23**	13.4	5.33	.32**	8.87	5.51	1.05**	9.07	4.97	1.03**
Male	15.4	5.47		15.1	5.32		15.35	6.34		15.2	6.13	
Black	12.26	4.93	.75**	12.18	4.93	.70**	8.83	5.59	1.11**	9.35	5.57	1.01**
White	16.23	5.33		15.8	5.22		15.62	6.23		15.4	6.06	
Asian	12.55	5.45	.67**	12.35	5.19	.65**	11.73	6.78	.60**	12.4	6.08	.49**
White	16.15	5.35		15.75	5.21		15.5	6.26		15.36	6.05	
White Hispanic	15.02	5.40	.23**	14.14	5.08	.34**	14	6.61	.26**	13.44	6.03	.35**
White Non-Hispanic	16.26	5.30		15.93	5.20		15.65	6.23		15.54	6.05	
Black Hispanic	12.9	5.09	.04	13.28	4.74	.16	10.11	6.43	.14	11.56	5.75	.33**
Black Non-Hispanic	12.7	4.91		12.48	4.98		9.26	5.85		9.68	5.76	

Subgroup	MK						BC					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	10.47	4.35	.34**	10.47	4.06	.39**	13.19	5.87	.56**	14.73	5.2	.46**
Male	11.98	4.52		12.14	4.42		16.46	5.85		17.2	5.36	
Black	9.6	3.80	.59**	10.06	3.70	.51**	11.07	5.12	1.10**	12.66	4.79	1.02**
White	12.17	4.48		12.25	4.37		17.12	5.59		17.85	5.15	
Asian	13.4	5.6	.28**	13.98	5.41	.40**	15.54	5.74	.26**	16.53	4.88	.24**
White	12.11	4.45		12.19	4.34		17.02	5.63		17.77	5.17	
White Hispanic	11.64	4.33		11.34	3.92		16.02	5.72		16.41	5.22	
White Non-Hispanic	12.14	4.48	.11*	12.28	4.39	.22**	17.11	5.60	.19**	17.88	5.13	.29**
Black Hispanic	9.98	3.51		10.65	3.57		12.16	5.60		14.23	4.98	
Black Non-Hispanic	9.84	4.04	.04	10.09	3.69	.15	11.35	5.32	.15	12.92	4.89	.27*
Subgroup	RC						AI					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	15.67	3.98	.50**	16.42	4.10	.26**	6.73	2.89	.88**	6.41	2.52	.84**
Male	17.55	3.73		17.45	3.96		9.84	3.69		9.26	3.56	
Black	14.95	3.80	.84**	15.37	4.03	.72**	6.97	3.02	.83**	6.78	2.80	.76**
White	17.97	3.57		18.05	3.69		9.97	3.69		9.41	3.56	
Asian	15.36	4.06	.71**	15.37	4.72	.71**	7.98	3.57	.53**	7.27	3.10	.60**
White	17.93	3.59		18.03	3.69		9.92	3.68		9.39	3.57	
White Hispanic	17.42	3.70		17.04	3.79		9.27	3.73		8.35	3.29	
White Non-Hispanic	17.96	3.58	.15**	18.12	3.67	.29**	9.96	3.67	.19**	9.48	3.55	.32**
Black Hispanic	15.7	4.40		16.06	3.94		7.41	3.43		7.01	2.62	
Black Non-Hispanic	15.23	3.77	.12	15.46	4.07	.15	7.13	3.09	.09	6.96	3.01	.02

[†] $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M = Mean; SD = Standard Deviation; d = Cohen's d . Underlined are Cohen's d for which minority test-takers outperformed majority test-takers.

Total T1: $N_{\text{Female}} = 1,114$, $N_{\text{Male}} = 4,308$;

Total T2: $N_{\text{Female}} = 940$, $N_{\text{Male}} = 3,900$;

Total T1: $N_{\text{Black}} = 671$, $N_{\text{White}} = 3,584$;

Total T2: $N_{\text{Black}} = 637$, $N_{\text{White}} = 3,151$;

Total T1: $N_{\text{Asian}} = 257$, $N_{\text{White}} = 3,584$;

Total T2: $N_{\text{Asian}} = 243$, $N_{\text{White}} = 3,188$;

Total T1: $N_{\text{White Non-Hispanic}} = 3,448$, $N_{\text{White Hispanic}} = 497$;

Total T2: $N_{\text{White Non-Hispanic}} = 3,024$, $N_{\text{White Hispanic}} = 420$;

Total T1: $N_{\text{Black Non-Hispanic}} = 867$, $N_{\text{Black Hispanic}} = 105$;

Total T2: $N_{\text{Black Non-Hispanic}} = 803$, $N_{\text{Black Hispanic}} = 96$.

For subtest-specific N , please refer to the Attachment 2.

AECF												
Subgroup	VA						PS					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	13.89	4.2	.02	11.97	4.35	.52**	6.63	2.71	.75**	7.84	2.76	.75**
Male	13.98	4.57		14.19	4.23		9.14	3.49		10.41	3.58	
Black	10.96	3.69	.97**	11.59	3.57	.76**	7.25	2.72	.54*	8.59	2.59	.43†
White	14.92	4.18		14.75	4.25		9.12	3.61		10.15	3.8	
Asian	13.71	4.82	.29	16.29	5.12	.39	8.43	2.76	.21	10.43	4.35	.05
White	14.94	4.22		14.6	4.28		9.17	3.54		10.24	3.8	
White Hispanic	13.95	3.85	.24	13.25	4.05	.31	8.85	3.92	.09	10.67	3.55	.16
White Non-Hispanic	14.95	4.28		14.55	4.18		9.18	3.51		10.07	3.78	
Black Hispanic	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Black Non-Hispanic	NA	NA		NA	NA		NA	NA		NA	NA	
Subgroup	AR						TR					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	10.11	4.92	.38*	9.39	4.17	.73**	28	6.56	.21	24.71	7.36	.34†
Male	11.92	4.68		12.81	4.84		26.55	7.12		27.02	6.5	
Black	7.71	3.96	1.05**	8.14	3.71	1.07**	24.04	6.51	.64**	22.05	5.96	.96**
White	12.54	4.71		13.1	4.78		28.39	6.86		28.09	6.35	
Asian	13.14	4.3	.12	14.14	5.18	.24	27.43	3.15	.12	29.43	5.62	.27
White	12.6	4.65		12.99	4.84		28.21	6.85		27.66	6.51	
White Hispanic	11.15	5.97	.30	14.25	4.9	.32	25.55	6.5	.48†	28.42	5.09	.10
White Non-Hispanic	12.56	4.41		12.71	4.8		28.79	6.82		27.77	6.67	
Black Hispanic	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Black Non-Hispanic	NA	NA		NA	NA		NA	NA		NA	NA	
Subgroup	WK						IC					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	12.26	5.6	.27	11.24	5.89	.56**	11.42	7.13	.46*	8.29	4.41	1.42**
Male	13.81	5.8		14.48	5.77		14.33	6.11		15.96	5.64	
Black	11.5	5.08	.53*	10.95	4.25	.73**	9.33	5.58	1.13**	9.27	5.58	1.12**
White	14.44	5.67		15.06	5.86		15.79	5.76		15.9	5.96	
Asian	12	5.42	.44	15.43	5	.05	10.43	4.31	.94*	12.86	5.08	.57
White	14.51	5.73		15.15	5.74		15.75	5.71		16.16	5.88	
White Hispanic	13.95	5	.08	12	3.98	.48	13.5	5.76	.40	14.58	6.61	.22
White Non-Hispanic	14.44	5.89		14.9	6.15		15.82	5.86		15.88	5.74	
Black Hispanic	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Black Non-Hispanic	NA	NA		NA	NA		NA	NA		NA	NA	

Subgroup	MK						BC					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	10.87	4.23	.33 [†]	10	3.84	.67**	13.92	6.37	.19	11.58	5.07	1.11**
Male	12.33	4.52		12.65	4		15.17	6.47		17.08	4.93	
Black	10.38	3.15	.50*	9.91	2.71	.71**	10.46	4.9	1.13**	10.86	4.25	1.22**
White	12.61	4.74		12.74	4.22		17.07	6.05		17.17	5.34	
Asian	14	3.16		13.43	5.88		12.43	5.44		16.43	2.37	
White	12.62	4.7	<u>.30</u>	12.51	4.18	<u>.21</u>	17.04	5.95	.78*	17.19	4.96	.16
White Hispanic	12.45	4.73	<u>.00</u>	13.08	4.85		14.9	6.43		18.08	5.48	
White Non-Hispanic	12.44	4.71		12.65	4.08	<u>.10</u>	17.04	6.03	.35	17.06	5.27	<u>.19</u>
Black Hispanic	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Black Non-Hispanic	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Subgroup	RC						AI					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	15.03	3.64	.39*	14.45	4.95	.59**	6.73	2.62	.44*	6.03	2.65	.75**
Male	16.49	3.82		17.05	4.19		8.23	3.63		8.46	3.41	
Black	14.58	3.63	.55*	13.36	3.44	.99**	5.67	2.79	.95**	5.45	2.44	1.07**
White	16.62	3.75		17.45	4.23		8.79	3.4		8.91	3.38	
Asian	16.57	2.57	.05	17	5.16	.11	6.86	2.54	.56	7.14	4.06	.57
White	16.75	3.67		17.48	4.22		8.74	3.43		9.1	3.37	
White Hispanic	17.35	3.28	<u>.21</u>	15.83	3.07	.38	7.35	3.96	.47 [†]	8	3.72	.21
White Non-Hispanic	16.56	3.79		17.42	4.3		8.93	3.2		8.67	3.23	
Black Hispanic	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Black Non-Hispanic	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

[†] $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M = Mean; SD = Standard Deviation; d = Cohen's d . NA means that the sample size was too small to generate stable statistics. Underlined are Cohen's d for which minority test-takers outperformed majority test-takers.

Total T1: $N_{\text{Female}} = 38$, $N_{\text{Male}} = 132$;

Total T2: $N_{\text{Female}} = 38$, $N_{\text{Male}} = 129$;

Total T1: $N_{\text{Black}} = 24$, $N_{\text{White}} = 108$;

Total T2: $N_{\text{Black}} = 22$, $N_{\text{White}} = 103$;

Total T1: $N_{\text{Asian}} = 7$, $N_{\text{White}} = 109$;

Total T2: $N_{\text{Asian}} = 7$, $N_{\text{White}} = 101$;

Total T1: $N_{\text{White Non-Hispanic}} = 96$, $N_{\text{White Hispanic}} = 20$;

Total T2: $N_{\text{White Non-Hispanic}} = 105$, $N_{\text{White Hispanic}} = 12$;

Total T1: $N_{\text{Black Non-Hispanic}} = 31$, $N_{\text{Black Hispanic}} = 5$;

Total T2: $N_{\text{Black Non-Hispanic}} = 33$, $N_{\text{Black Hispanic}} = 2$.

For subtest-specific N , please refer to the Attachment 2.

USAFA												
Subgroup	VA						PS					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	15.69	3.82	.05	15.96	4.09	.03	11.73	3.09	.66**	12.25	2.92	.66**
Male	15.89	3.81		16.06	3.91		13.78	3.1		14.22	3.03	
Black	13.46	3.88	.73**	14.06	4.11	.61**	11.13	3.43	.76**	12.02	3.13	.63**
White	16.15	3.69		16.41	3.81		13.53	3.13		13.95	3.08	
Asian	15.47	4.1	.18**	16	4.14	.09	12.89	3.5	.19**	13.76	2.91	.05
White	16.14	3.69		16.35	3.85		13.51	3.12		13.9	3.09	
White Hispanic	15.71	3.75	.12*	15.52	3.89	.23**	12.91	3.21	.21**	13.58	3.15	.11†
White Non-Hispanic	16.15	3.7		16.39	3.81		13.55	3.1		13.93	3.08	
Black Hispanic	14.3	3.49	.09	13.5	3.95	.14	11.79	3.54	.07	13.19	3.1	.36*
Black Non-Hispanic	13.94	4.01		14.07	4.2		11.57	3.32		12.06	3.18	
Subgroup	AR						TR					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	14.47	4.35	.51**	14.46	4.56	.55**	28.36	5.51	.02	27.64	5.75	.02
Male	16.67	4.33		16.93	4.46		28.46	5.55		27.54	5.8	
Black	12.51	4.23	.93**	13.16	4.7	.79**	24.86	5.25	.73**	24.85	5.49	.56**
White	16.5	4.28		16.69	4.46		28.82	5.45		28.02	5.63	
Asian	16.62	4.8	.05	17.44	4.62	.19**	28.58	5.29	.03	27.97	5.76	.01
White	16.42	4.3		16.59	4.47		28.77	5.4		27.93	5.69	
White Hispanic	15.4	4.65	.26**	15.69	4.41	.22**	27.86	5.4	.18**	27.72	5.8	.03
White Non-Hispanic	16.52	4.25		16.65	4.46		28.86	5.45		27.9	5.69	
Black Hispanic	13.06	4.28	.02	14.67	5.2	.31†	26.7	5.74	.26†	25.1	5.68	.04
Black Non-Hispanic	13.13	4.3		13.24	4.54		25.3	5.33		24.84	5.83	
Subgroup	WK						IC					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	16.07	4.96	.03	15.32	4.89	.13**	11.1	5.48	1.12**	10.89	5.18	1.10**
Male	16.24	5.07		15.94	4.95		17.46	5.73		17.03	5.72	
Black	12.99	5.03	.73**	13.54	5.08	.56**	11.3	6.32	.83**	11.2	6.04	.81**
White	16.6	4.92		16.22	4.78		16.4	6.11		16.08	5.99	
Asian	15.97	4.93	.12†	15.39	5.07	.16*	14.23	6.54	.35**	14.12	6.39	.32**
White	16.57	4.94		16.15	4.82		16.39	6.13		16.05	6.01	
White Hispanic	15.92	4.99	.14**	15.23	4.74	.20**	15.16	6.19	.22**	15.34	6.12	.12†
White Non-Hispanic	16.61	4.92		16.2	4.82		16.51	6.09		16.06	6.02	
Black Hispanic	13.32	5.88	.09	13.26	5.18	.01	12.3	6.93	.04	12.26	6.1	.04
Black Non-Hispanic	13.75	4.96		13.31	5.16		12.53	6.54		12.02	6.39	

Subgroup	MK						BC					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	17.43	4.07		17.11	3.81		11.93	4.68		13.45	4.4	
Male	18.79	4.05	.34**	18.43	3.87	.34**	14.34	5.26	.47**	15.52	4.66	.45**
Black	15.35	4.08		15.64	3.97		10.36	4.66		11.81	4.71	
White	18.73	3.96	.85**	18.39	3.77	.73**	14.07	5.16	.72**	15.35	4.6	.77**
Asian	19.4	4.06		19.69	3.61		13.59	5.38		15.21	4.47	
White	18.66	3.98	.19**	18.28	3.79	.37**	14.03	5.16	.09	15.34	4.63	.03
White Hispanic	17.81	4.15		17.6	3.92		12.94	5		14.71	4.46	
White Non-Hispanic	18.74	3.96	.23**	18.33	3.76	.19**	14.15	5.17	.24**	15.32	4.62	.13*
Black Hispanic	15.98	4.63		16.62	3.91		11.74	5.57		12.69	4.62	
Black Non-Hispanic	15.84	4.08	.03	15.49	3.94	.29†	11.2	4.84	.11	12.58	4.87	.02

Subgroup	RC						AI					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	17.39	3.61		18.37	3.84		8.79	3.15		8.37	2.68	
Male	18.4	3.69	.27**	18.87	3.65	.13**	11.48	3.86	.73**	10.82	3.51	.74**
Black	15.36	4.44		16.63	4.26		8.35	3.39		8.36	3.15	
White	18.5	3.45	.89**	19.1	3.48	.70**	11.19	3.8	.75**	10.51	3.48	.62**
Asian	17.69	3.81		18.08	3.95		9.34	3.49		9.09	3.24	
White	18.49	3.48	.23**	19.07	3.53	.28**	11.22	3.82	.50**	10.52	3.5	.41**
White Hispanic	18.14	3.62		18.56	3.55		10.29	3.7		9.81	3.48	
White Non-Hispanic	18.48	3.47	.10†	19.07	3.52	.14*	11.26	3.83	.25**	10.53	3.47	.21**
Black Hispanic	15.57	4.14		16.79	4.52		7.41	3.43		7.01	2.62	
Black Non-Hispanic	15.98	4.47	.09	16.47	4.54	.07	7.13	3.09	.09	6.96	3.01	.02

† $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M = Mean; SD = Standard Deviation; d = Cohen's d . Underlined are Cohen's d for which minority test-takers outperformed majority test-takers.

Total T1: $N_{\text{Female}} = 1,114$, $N_{\text{Male}} = 3,195$;

Total T2: $N_{\text{Female}} = 945$, $N_{\text{Male}} = 2,791$;

Total T1: $N_{\text{Black}} = 261$, $N_{\text{White}} = 3,427$;

Total T2: $N_{\text{Black}} = 184$, $N_{\text{White}} = 2,879$;

Total T1: $N_{\text{Asian}} = 277$, $N_{\text{White}} = 3,357$;

Total T2: $N_{\text{Asian}} = 212$, $N_{\text{White}} = 2,801$;

Total T1: $N_{\text{White Non-Hispanic}} = 3,224$, $N_{\text{White Hispanic}} = 384$;

Total T2: $N_{\text{White Non-Hispanic}} = 2,843$, $N_{\text{White Hispanic}} = 275$;

Total T1: $N_{\text{Black Non-Hispanic}} = 367$, $N_{\text{Black Hispanic}} = 47$;

Total T2: $N_{\text{Black Non-Hispanic}} = 287$, $N_{\text{Black Hispanic}} = 42$.

For subtest-specific N , please refer to the Attachment 2.

ROTC												
Subgroup	VA						PS					
	M	T1 SD	d	M	T2 SD	d	M	T1 SD	d	M	T2 SD	d
Female	13.43	4.59	.23**	13.27	4.81	.21**	9.03	3.51	.70**	9.39	3.5	.70**
Male	14.45	4.31		14.23	4.55		11.65	3.84		12.02	3.82	
Black	11.26	4.17	.89**	10.91	4.39	.90**	8.13	3.28	.92**	8.31	3.33	.98**
White	14.95	4.15		14.79	4.33		11.6	3.81		11.94	3.75	
Asian	13.88	4.39	.24**	13.53	4.75	.27**	10.86	3.78	.18**	11.19	3.73	.19**
White	14.89	4.16		14.7	4.35		11.54	3.83		11.91	3.76	
White Hispanic	13.21	4.4	.46**	13.12	4.59	.42**	10.17	3.75	.40**	10.84	3.74	.31**
White Non-Hispanic	15.11	4.09		14.93	4.29		11.7	3.8		12.01	3.76	
Black Hispanic	11.72	4.35	.07	11.41	4.47	.09	8.4	3.8	.04	9.11	3.53	.18*
Black Non-Hispanic	11.42	4.21		11.01	4.48		8.27	3.37		8.48	3.53	
Subgroup	AR						TR					
	M	T1 SD	d	M	T2 SD	d	M	T1 SD	d	M	T2 SD	d
Female	10.9	4.53	.57**	10.85	4.73	.56**	25.6	6.08	.11**	25.02	6.33	.12**
Male	13.69	4.99		13.67	5.11		26.26	6.15		25.78	6.36	
Black	9.35	4.11	.92**	9.02	4.1	.96**	22.06	5.59	.83**	21.52	5.77	.80**
White	13.74	4.84		13.68	4.96		26.97	5.93		26.4	6.12	
Asian	13.7	5.21	.01	13.68	5.47	.02	26.44	5.99	.08*	26.19	6.69	.02
White	13.65	4.86		13.6	4.97		26.89	5.94		26.33	6.13	
White Hispanic	11.78	4.7	.44**	11.85	4.8	.40**	25.02	6.24	.36**	24.67	6.09	.31**
White Non-Hispanic	13.89	4.82		13.82	4.96		27.15	5.85		26.55	6.1	
Black Hispanic	9.57	4.64	.03	9.74	4.73	.10	22.73	6.05	.05	22.42	6.1	.09
Black Non-Hispanic	9.46	4.08		9.29	4.28		22.44	5.74		21.89	5.87	
Subgroup	WK						IC					
	M	T1 SD	d	M	T2 SD	d	M	T1 SD	d	M	T2 SD	d
Female	12.93	5.42	.20**	12.33	5.33	.30**	9.3	5.57	1.18**	9.74	5.24	1.11**
Male	14.01	5.56		13.95	5.31		16.39	6.21		16.21	6.05	
Black	10.71	4.84	.71**	10.15	4.62	.84**	8.52	5.68	1.15**	8.8	5.43	1.14**
White	14.54	5.45		14.44	5.16		15.79	6.42		15.68	6.12	
Asian	13.18	5.36	.24**	12.41	5.27	.38**	12.8	6.52	.45**	13.03	6.18	.42**
White	14.47	5.45		14.36	5.17		15.71	6.45		15.62	6.16	
White Hispanic	12.58	5.41	.40**	12.82	5.3	.34**	13.23	6.65	.44**	13.57	6.11	.37**
White Non-Hispanic	14.74	5.41		14.56	5.13		16.02	6.34		15.84	6.11	
Black Hispanic	11.67	5.22	.16*	10.72	4.8	.09	10.44	6.42	.27**	10.82	6.12	.29**
Black Non-Hispanic	10.87	4.87		10.27	4.8		8.83	5.9		9.13	5.67	

Subgroup	MK						BC					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	13.04	5.1	.45**	13.03	4.89	.44**	12.15	5.46	.54**	13.51	4.97	.50**
Male	15.38	5.23		15.23	4.97		15.18	5.59		15.98	4.95	
Black	11.09	4.45	.88**	10.88	4.04	.94**	9.38	4.77	1.10**	11.1	4.49	1.05**
White	15.5	5.1		15.36	4.83		15.32	5.48		16.13	4.8	
Asian	16.45	5.21	<u>.21**</u>	15.95	5.12	<u>.14**</u>	14.46	5.27	.14**	15.38	4.85	.14**
White	15.39	5.13		15.26	4.85		15.24	5.5		16.07	4.82	
White Hispanic	13.54	5.07	.41**	13.68	4.75	.37**	13.53	5.56	.35**	14.6	4.7	.35**
White Non-Hispanic	15.62	5.1		15.45	4.84		15.45	5.45		16.25	4.8	
Black Hispanic	11.08	5.13	.01	11.69	4.39	<u>.15*</u>	10.8	5.18	<u>.23**</u>	12.93	5.16	<u>.33**</u>
Black Non-Hispanic	11.1	4.51		11.04	4.26		9.69	4.86		11.37	4.69	
Subgroup	RC						AI					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	15.32	4.22	.43**	15.96	4.87	.35**	6.33	2.96	.83**	6.19	2.59	.85**
Male	17.07	4.02		17.52	4.33		9.53	4.15		9.16	3.81	
Black	13.78	4.14	.91**	13.44	4.69	1.15**	5.87	2.93	.87**	5.81	2.71	.87**
White	17.32	3.86		18.09	3.95		9.36	4.11		8.98	3.75	
Asian	16.07	4.07	.31**	15.69	4.89	.57**	7.35	3.46	.49**	7.04	3.28	.52**
White	17.26	3.89		18.03	3.97		9.33	4.12		8.98	3.75	
White Hispanic	15.9	4.2	.40**	16.47	4.58	.44**	7.7	3.74	.45**	7.68	3.43	.38**
White Non-Hispanic	17.43	3.81		18.21	3.87		9.51	4.12		9.09	3.76	
Black Hispanic	14.3	4.4	<u>.09</u>	14.46	5.12	<u>.18*</u>	6.33	3.36	<u>.09</u>	6.5	2.92	<u>.19**</u>
Black Non-Hispanic	13.92	4.1		13.62	4.73		6.04	3.1		5.94	2.86	

[†] $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M = Mean; SD = Standard Deviation; d = Cohen's d . Underlined are Cohen's d for which minority test-takers outperformed majority test-takers.

Total T1: $N_{\text{Female}} = 3,604$, $N_{\text{Male}} = 9,148$;

Total T2: $N_{\text{Female}} = 3,552$, $N_{\text{Male}} = 8,834$;

Total T1: $N_{\text{Black}} = 1,093$, $N_{\text{White}} = 8,972$;

Total T2: $N_{\text{Black}} = 1,068$, $N_{\text{White}} = 8,753$;

Total T1: $N_{\text{Asian}} = 880$, $N_{\text{White}} = 8,760$;

Total T2: $N_{\text{Asian}} = 863$, $N_{\text{White}} = 8,584$;

Total T1: $N_{\text{White Non-Hispanic}} = 8,222$, $N_{\text{White Hispanic}} = 1,317$;

Total T2: $N_{\text{White Non-Hispanic}} = 7,978$, $N_{\text{White Hispanic}} = 1,310$;

Total T1: $N_{\text{Black Non-Hispanic}} = 1,392$, $N_{\text{Black Hispanic}} = 199$;

Total T2: $N_{\text{Black Non-Hispanic}} = 1,337$, $N_{\text{Black Hispanic}} = 206$.

For subtest-specific N , please refer to the Attachment 2.

ANG												
Subgroup	VA						PS					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	14.42	4.23	.15**	14.19	4.36	.19**	7.82	3.08	.65**	8.36	3.13	.70**
Male	15.06	4.21		15.01	4.33		10.16	3.71		10.78	3.54	
Black	12.38	4.27	.77**	12.24	4.42	.76**	7.51	3.31	.69**	7.96	3.28	.78**
White	15.49	3.99		15.41	4.16		10.03	3.7		10.68	3.52	
Asian	13.36	4.77	.52**	13.77	4.4	.38**	9.14	3.49	.24**	10.99	3.41	.10
White	15.47	4.01		15.37	4.17		10.01	3.69		10.65	3.52	
White Hispanic	14.15	4	.34**	14.47	4.52	.23**	8.98	3.41	.30**	9.8	3.44	.25**
White Non-Hispanic	15.53	4.01		15.42	4.14		10.08	3.71		10.68	3.52	
Black Hispanic	12.66	4.16	.05	12.55	4.31	.02	8.44	3.06	.19	8.55	3	.20
Black Non-Hispanic	12.43	4.37		12.46	4.47		7.79	3.38		7.92	3.26	
Subgroup	AR						TR					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	11.12	4.42	.48**	10.26	4.25	.62**	27.89	6.42	.13**	27.46	6.75	.14**
Male	13.39	4.76		13.19	4.84		28.7	6.37		28.37	6.6	
Black	9.84	4.11	.79**	9.22	4.04	.83**	24.55	6.68	.74**	22.74	6.69	.98**
White	13.47	4.67		13.12	4.75		29.2	6.22		28.99	6.34	
Asian	13.37	5	.01	13.41	5.04	.07	27.95	5.8	.19*	27.94	5.76	.16†
White	13.43	4.69		13.09	4.75		29.11	6.23		28.97	6.34	
White Hispanic	11.75	4.45	.39**	11.47	4.64	.36**	27.2	6.36	.34**	27.31	6.17	.27**
White Non-Hispanic	13.55	4.67		13.19	4.74		29.28	6.15		29.05	6.38	
Black Hispanic	9.75	4.03	.04	9	4.29	.10	25.81	6.17	.18	23.24	6.57	.05
Black Non-Hispanic	9.94	4.22		9.42	4.07		24.62	6.58		23.61	7.01	
Subgroup	WK						IC					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	14.57	5.52	.19**	13.84	5.39	.32**	10.22	5.97	1.00**	9.99	5.52	1.07**
Male	15.62	5.53		15.55	5.32		16.56	6.4		16.47	6.19	
Black	12.53	5.17	.63**	12.25	5.22	.69**	9.81	6.25	.97**	9.22	5.26	1.12**
White	15.95	5.42		15.85	5.23		16.15	6.56		16.17	6.31	
Asian	13.14	5.52	.51**	13.49	5.62	.44**	13.63	6.19	.38**	12.75	6.5	.54**
White	15.92	5.42		15.81	5.24		16.1	6.57		16.17	6.29	
White Hispanic	14.7	5.32	.24**	15.16	5.1	.13†	14.21	6.45	.30**	14.12	6.77	.34**
White Non-Hispanic	16.01	5.44		15.82	5.25		16.2	6.55		16.26	6.24	
Black Hispanic	13.38	5.64	.11	13.72	4.41	.24	10.88	6.34	.17	11.24	5.41	.34†
Black Non-Hispanic	12.78	5.36		12.5	5.28		9.82	6.29		9.4	5.45	

Subgroup	MK						BC					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	11.18	4.44	.32**	10.88	4	.47**	14.2	6.03	.47**	14.75	5.27	.57**
Male	12.71	4.89		12.97	4.56		16.95	5.83		17.69	5.15	
Black	10.32	4.16	.51**	9.63	3.51	.76**	11.7	5.52	.94**	11.98	4.99	1.18**
White	12.75	4.81		12.96	4.47		17.12	5.8		17.84	4.96	
Asian	14.27	5.43	.32**	13.91	5.03	.22*	16.26	5.35	.14	16.71	5.56	.22*
White	12.71	4.79		12.9	4.47		17.07	5.79		17.81	4.97	
White Hispanic	11.74	4.63	.21**	11.6	4.37	.32**	15.1	5.7	.36**	16.29	5.22	.33**
White Non-Hispanic	12.74	4.79		13.01	4.47		17.2	5.78		17.92	4.92	
Black Hispanic	11.16	4.59	.20	9.34	3.7	.15	13.09	5.72	.23	12.93	5.44	.11
Black Non-Hispanic	10.31	4.15		9.88	3.61		11.84	5.48		12.37	5.11	

Subgroup	RC						AI					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	16.3	3.86	.40**	17.05	4.09	.26**	7.85	3.83	.85**	7.17	3.39	.93**
Male	17.82	3.75		18.05	3.87		11.74	4.78		11.15	4.48	
Black	15.02	3.93	.82**	15.19	4.28	.88**	7.79	4.08	.78**	7	3.62	.92**
White	17.98	3.58		18.43	3.63		11.49	4.8		11.03	4.47	
Asian	15.58	4.41	.66**	16.27	4.44	.58**	9.22	4.63	.46**	8.61	4.02	.54**
White	17.97	3.58		18.41	3.63		11.44	4.81		11.01	4.47	
White Hispanic	16.88	3.64	.33**	17.26	3.98	.33**	9.81	4.6	.36**	9.43	4.5	.36**
White Non-Hispanic	18.05	3.57		18.46	3.61		11.54	4.79		11.05	4.45	
Black Hispanic	15.16	3.54	.03	15.28	4.28	.02	8.5	3.65	.15	7.79	4.53	.19
Black Non-Hispanic	15.29	4.07		15.35	4.3		7.9	4.11		7.11	3.6	

[†] $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M = Mean; SD = Standard Deviation; d = Cohen's d . Underlined are Cohen's d for which minority test-takers outperformed majority test-takers.

Total T1: $N_{\text{Female}} = 863$, $N_{\text{Male}} = 3,050$;

Total T2: $N_{\text{Female}} = 691$, $N_{\text{Male}} = 2,595$;

Total T1: $N_{\text{Black}} = 284$, $N_{\text{White}} = 2,941$;

Total T2: $N_{\text{Black}} = 246$, $N_{\text{White}} = 2,465$;

Total T1: $N_{\text{Asian}} = 129$, $N_{\text{White}} = 2,902$;

Total T2: $N_{\text{Asian}} = 111$, $N_{\text{White}} = 2,434$;

Total T1: $N_{\text{White Non-Hispanic}} = 2,896$, $N_{\text{White Hispanic}} = 306$;

Total T2: $N_{\text{White Non-Hispanic}} = 2,432$, $N_{\text{White Hispanic}} = 232$;

Total T1: $N_{\text{Black Non-Hispanic}} = 389$, $N_{\text{Black Hispanic}} = 32$;

Total T2: $N_{\text{Black Non-Hispanic}} = 334$, $N_{\text{Black Hispanic}} = 29$.

For subtest-specific N , please refer to the Attachment 2.

AFRES												
Subgroup	VA						PS					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	13.81	4.39	.23**	13.79	4.69	.17**	7.5	3.24	.68**	8.24	3.35	.67**
Male	14.82	4.32		14.53	4.3		9.98	3.77		10.63	3.65	
Black	12.23	4.05	.84**	12.18	4.22	.71**	7.28	3.08	.77**	7.79	3.47	.82**
White	15.63	4.06		15.17	4.18		10.06	3.73		10.69	3.58	
Asian	13.58	4.35	.48**	14.29	4.6	.20	9.27	3.71	.21†	10.47	4.24	.04
White	15.55	4.1		15.13	4.21		10.03	3.73		10.63	3.62	
White Hispanic	14.15	4.25	.38**	13.65	4.38	.38**	8.99	3.69	.30**	9.8	3.54	.25*
White Non-Hispanic	15.69	4.04		15.26	4.18		10.11	3.73		10.7	3.57	
Black Hispanic	10.69	3.64	.41*	13.06	3.87	<u>.21</u>	6.55	3.16		8.25	3.15	
Black Non-Hispanic	12.3	3.99		12.16	4.39		7.49	3.11	.30	7.62	3.47	<u>.18</u>
Subgroup	AR						TR					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	10.85	4.46	.46**	10.26	4.85	.58**	27.66	6.38	.05	27.14	6.31	.08
Male	12.99	4.77		13.12	4.99		27.95	6.51		27.63	6.63	
Black	9.74	3.82	.85**	9.48	4.44	.78**	24.2	6.21	.81**	23.23	5.91	.87**
White	13.58	4.67		13.24	4.93		29.18	6.15		28.69	6.33	
Asian	12.55	4.38	.21†	13.42	5.64	<u>.05</u>	28.66	5.88	.0	27.56	6.03	.17
White	13.51	4.71		13.18	4.95		29.13	6.15		28.61	6.32	
White Hispanic	11.88	5.04	.37**	11.21	4.96	.44**	27.65	6.5	.27**	26.96	5.97	.29**
White Non-Hispanic	13.59	4.6		13.38	4.92		29.31	6.05		28.81	6.34	
Black Hispanic	8.86	3.38	.24	10.38	4.65	<u>.21</u>	25.62	4.91	<u>.26</u>	26.06	6.98	
Black Non-Hispanic	9.77	3.89		9.42	4.45		24.02	6.31		23.52	6.09	<u>.41</u>
Subgroup	WK						IC					
	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T1</i> <i>SD</i>	<i>d</i>	<i>M</i>	<i>T2</i> <i>SD</i>	<i>d</i>
Female	14.32	5.7	.19**	14.26	5.71	.20**	9.23	5.8	1.05**	10.12	5.78	1.01**
Male	15.4	5.8		15.38	5.51		15.97	6.6		16.38	6.37	
Black	11.74	5.12	.87**	11.77	4.94	.87**	8.79	5.73	1.19**	8.57	5.43	1.33**
White	16.47	5.5		16.32	5.31		16.3	6.42		16.51	6.09	
Asian	12.86	6.06	.64**	14.89	6.06	.25†	12.58	6.17	.56**	13.55	6.64	.45**
White	16.42	5.54		16.22	5.36		16.19	6.52		16.35	6.16	
White Hispanic	14.79	5.38	.33**	14.96	5.49	.24*	13.69	7.17	.42**	14.41	6.6	.36**
White Non-Hispanic	16.61	5.5		16.23	5.35		16.45	6.41		16.6	6.09	
Black Hispanic	11.62	4.07	.09	11.5	4.77	.07	9.34	5.01	<u>.10</u>	10.19	7.02	
Black Non-Hispanic	12.07	5.05		11.85	5.08		8.77	6.02		8.76	5.62	<u>.25</u>

Subgroup	MK						BC					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	10.94	4.41	.37**	11.07	4.68	.44**	13.24	6.03	.49**	14.58	5.7	.51**
Male	12.73	5.01		13.14	4.72		16.33	6.35		17.39	5.39	
Black	10.06	3.97	.63**	10.69	4.55	.52**	11.04	5.25	1.06**	11.86	4.91	1.18**
White	13.08	4.98		13.17	4.78		17.24	5.94		17.87	5.16	
Asian	13.8	4.79	<u>.16</u>	14.61	5.41	<u>.31*</u>	16.03	6.05	.19	16.32	5.59	.29*
White	12.99	5.01		13.1	4.78		17.17	5.97		17.85	5.19	
White Hispanic	11.79	5.05	.25**	11.56	4.48	.36**	14.6	6.23	.47**	16.24	5.24	.34**
White Non-Hispanic	13.02	4.92		13.27	4.73		17.39	5.9		18.01	5.15	
Black Hispanic	8.97	4.01	.28	10.38	3.12	.04	11.03	5.97	.01	15.13	4.91	.57*
Black Non-Hispanic	10.1	4.07		10.54	4.28		11.08	5.24		12.17	5.24	<u>.57*</u>
Subgroup	RC						AI					
	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>	<i>M</i>	T1 <i>SD</i>	<i>d</i>	<i>M</i>	T2 <i>SD</i>	<i>d</i>
Female	15.76	3.89	.47**	16.8	4.22	.20**	7.55	3.94	.89**	7.08	3.59	.92**
Male	17.61	3.94		17.62	3.97		11.75	4.97		10.98	4.45	
Black	14.52	4.04	1.02**	15.11	4.08	.85**	7.27	3.78	.99**	6.48	3.19	1.05**
White	18.19	3.47		18.31	3.69		11.91	4.88		10.94	4.46	
Asian	16.31	4.18	.51**	16.66	3.93	.43**	9.1	4.09	.58**	8.97	4.41	.42**
White	18.1	3.5		18.25	3.71		11.88	4.88		10.84	4.42	
White Hispanic	16.82	3.48	.40**	16.88	3.93	.40**	9.71	4.83	.50**	9.31	4.47	.41**
White Non-Hispanic	18.24	3.55		18.37	3.65		12.15	4.86		11.09	4.38	
Black Hispanic	13.69	4.00	.26	15.88	3.52	<u>.18</u>	7.28	3.7	.06	6.94	2.24	<u>.08</u>
Black Non-Hispanic	14.67	3.81		15.14	4.07		7.52	3.94		6.68	3.47	

[†] $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M = Mean; SD = Standard Deviation; d = Cohen's d . Underlined are Cohen's d for which minority test-takers outperformed majority test-takers.

Total T1: $N_{\text{Female}} = 367$, $N_{\text{Male}} = 1,111$;

Total T2: $N_{\text{Female}} = 337$, $N_{\text{Male}} = 877$;

Total T1: $N_{\text{Black}} = 203$, $N_{\text{White}} = 931$;

Total T2: $N_{\text{Black}} = 162$, $N_{\text{White}} = 785$;

Total T1: $N_{\text{Asian}} = 71$, $N_{\text{White}} = 918$;

Total T2: $N_{\text{Asian}} = 62$, $N_{\text{White}} = 794$;

Total T1: $N_{\text{White Non-Hispanic}} = 871$, $N_{\text{White Hispanic}} = 150$;

Total T2: $N_{\text{White Non-Hispanic}} = 740$, $N_{\text{White Hispanic}} = 111$;

Total T1: $N_{\text{Black Non-Hispanic}} = 263$, $N_{\text{Black Hispanic}} = 29$;

Total T2: $N_{\text{Black Non-Hispanic}} = 211$, $N_{\text{Black Hispanic}} = 16$.

For subtest-specific N , please refer to the Attachment 2.

APPENDIX L. Measurement Invariance Chi Square Tests

Female/Male**											
Model	T1						T2				
	AIC	BIC	χ^2	χ^2 diff	df	df diff	AIC	BIC	χ^2	χ^2 diff	df
Configural	5,452,201	5,455,904	62,623		2,424		4,967,074	4,970,737	52,030		2,424
Metric	5,454,383	5,457,169	65,021	2,397.7	2,532	108	4,969,315	4,972,071	54,487	2,457.4	2,532
Scalar	5,456,656	5,459,051	67,386	2,365.3	2,578	46	4,972,542	4,974,911	57,806	3,318.7	2,578
Black/White**											
Model	T1						T2				
	AIC	BIC	χ^2	χ^2 diff	df	df diff	AIC	BIC	χ^2	χ^2 diff	df
Configural	4,331,601	4,335,204	50,417		2,424		3,932,336	3,935,898	42,539		2,424
Metric	4,333,571	4,336,282	52,603	2,186.1	2,532	108	3,934,310	3,936,989	44,729	2,189.75	2,532
Scalar	4,334,741	4,337,072	53,865	1,262.3	2,578	46	3,935,019	3,937,323	45,530	801.46	2,578
Asian/White**											
Model	T1						T2				
	AIC	BIC	χ^2	χ^2 diff	df	df diff	AIC	BIC	χ^2	χ^2 diff	df
Configural	4,097,026	4,100,606	48,745		2,424		3,726,791	3,730,330	41,266		2,424
Metric	4,097,393	4,100,085	49,327	582.11	2,532	108	3,727,208	3,729,870	41,899	633.42	2,532
Scalar	4,098,515	4,100,830	50,542	1,214.75	2,578	46	3,728,410	3,730,698	43,192	1,293.23	2,578

**All chi square/df tests were significant at $p \leq .010$.

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; χ^2 = Chi Square; χ^2 diff = Chi Square Difference between the Models; df = Degrees of Freedom; df diff = Degrees of Freedom Difference between the Models.

APPENDIX M. Stratified Samples Internal Consistency

Subtest	T1 (Cronbach's Alpha)						
	OTS-CIV	OTS-AD	AACP	USAFA	ROTC	ANG	AFRES
VA	.77	.71	.74	.67	.73	.71	.73
AR	.81	.77	.78	.77	.80	.78	.78
WK	.86	.85	.86	.82	.84	.85	.87
MK	.83	.75	.75	.77	.83	.79	.80
RC	.74	.71	.68	.70	.74	.71	.73
PS	.76	.66	.66	.69	.75	.70	.71
TR	.89	.89	.90	.87	.87	.89	.89
IC	.92	.91	.90	.90	.91	.92	.92
BC	.87	.86	.88	.81	.84	.86	.88
AI	.81	.73	.69	.75	.78	.85	.87
Subtest	T2 (Cronbach's Alpha)						
	OTS-CIV	OTS-AD	AACP	USAFA	ROTC	ANG	AFRES
VA	.77	.72	.72	.69	.76	.73	.73
AR	.83	.78	.79	.79	.81	.79	.81
WK	.86	.84	.87	.81	.83	.84	.85
MK	.82	.74	.70	.74	.81	.76	.78
RC	.79	.74	.78	.73	.79	.74	.75
PS	.74	.63	.67	.66	.75	.68	.70
TR	.89	.89	.89	.87	.88	.89	.89
IC	.91	.90	.89	.90	.90	.91	.92
BC	.84	.83	.83	.77	.81	.82	.84
AI	.76	.69	.68	.67	.72	.82	.82

Note. For subtest-specific *N*, please refer to the Attachment 2.

APPENDIX N. Stratified Test-Retest Reliabilities

OTS-CIV										
Subtest	T1-T1 (same form)					T2-T2 (same form)				
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T1-T1}}$	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T2-T2}}$
VA	13.39	4.73	14.09	4.68	.87**	14.71	4.52	14.81	4.42	.94**
AR	12.41	5.05	13.17	4.82	.87**	13.36	5.01	13.47	5.06	.90**
WK	12.84	5.91	13.79	6.00	.90**	14.63	5.76	14.81	5.80	.93**
MK	12.61	5.17	13.79	5.16	.85**	13.19	5.19	13.73	4.99	.92**
RC	15.41	4.44	16.28	4.16	.79**	17.26	4.07	17.41	3.92	.95**
PS	9.69	3.82	9.97	3.86	.92**	10.51	3.09	10.55	3.12	.93**
TR	26.98	6.73	28.23	6.67	.86**	26.44	6.45	26.88	7.28	.90**
IC	13.79	7.15	15.44	7.08	.84**	14.73	6.60	14.86	6.83	.94**
BC	15.65	6.43	16.91	6.46	.85**	16.86	5.35	17.37	5.25	.91**
AI	9.18	4.76	9.80	4.85	.94**	8.16	4.21	8.45	4.39	.94**
OTS-AD										
Subtest	T1-T1 (same form)					T2-T2 (same form)				
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T1-T1}}$	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T2-T2}}$
VA	13.58	3.71	14.43	3.59	.82**	12.34	4.56	13.41	4.08	.76**
AR	11.17	4.06	11.95	4.30	.79**	10.08	4.29	11.11	4.40	.84**
WK	13.05	5.55	13.95	5.37	.90**	13.23	5.58	13.65	5.47	.92**
MK	10.37	3.86	11.78	3.56	.70**	10.53	3.71	11.33	3.65	.72**
RC	16.11	4.23	17.06	3.81	.82**	15.08	4.69	15.66	4.56	.89**
PS	8.36	3.21	8.41	3.46	.87**	8.45	2.60	8.37	2.85	.78**
TR	28.09	6.34	29.08	6.56	.67**	26.27	6.96	27.80	6.36	.69**
IC	14.10	6.15	15.35	6.32	.85**	11.28	5.61	12.30	6.15	.75**
BC	15.66	5.91	16.63	6.10	.76**	14.13	4.87	15.13	4.94	.81**
AI	8.77	3.90	9.47	4.22	.89**	7.64	3.09	8.18	3.39	.79**
AECP										
Subtest	T1-T1 (same form)					T2-T2 (same form)				
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T1-T1}}$	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T2-T2}}$
VA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
WK	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
MK	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RC	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
TR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
IC	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
BC	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AI	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
USAFA										
Subtest	T1-T1 (same form)					T2-T2 (same form)				
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T1-T1}}$	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T2-T2}}$
VA	15.11	3.46	15.48	3.51	.90**	NA	NA	NA	NA	NA
AR	15.21	4.05	15.23	4.13	.93**	NA	NA	NA	NA	NA
WK	14.38	4.25	14.38	4.52	.97**	NA	NA	NA	NA	NA
MK	17.38	4.03	17.56	3.93	.95**	NA	NA	NA	NA	NA
RC	16.91	4.32	17.01	4.17	.98**	NA	NA	NA	NA	NA
PS	12.74	3.06	12.95	2.95	.96**	NA	NA	NA	NA	NA
TR	26.78	5.46	27.04	5.62	.96**	NA	NA	NA	NA	NA
IC	13.79	6.78	14.55	6.77	.96**	NA	NA	NA	NA	NA
BC	12.40	5.12	12.77	5.14	.95**	NA	NA	NA	NA	NA
AI	10.06	3.62	10.59	3.60	.93**	NA	NA	NA	NA	NA

ROTC										
Subtest	T1-T1 (same form)					T2-T2 (same form)				
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T1-T1}}$	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T2-T2}}$
VA	10.59	3.77	12.72	3.65	.70**	10.37	4.08	12.79	4.01	.70**
AR	9.51	3.89	10.50	4.03	.72**	9.06	4.03	10.09	4.03	.72**
WK	9.51	4.44	10.89	4.64	.77**	9.83	4.74	11.05	4.73	.80**
MK	11.26	4.33	12.40	4.15	.74**	10.66	4.13	11.90	3.76	.71**
RC	13.62	3.85	15.18	3.87	.70**	13.57	4.55	15.22	4.46	.70**
PS	8.49	3.10	8.70	3.16	.73**	8.82	3.44	9.11	3.42	.80**
TR	24.63	6.13	26.67	6.69	.67**	23.32	6.09	24.96	7.21	.57**
IC	11.40	6.83	13.95	7.16	.78**	11.68	6.15	13.37	6.54	.82**
BC	12.44	5.45	14.47	5.89	.77**	13.22	4.56	15.17	4.94	.75**
AI	7.41	4.15	8.38	4.45	.83**	7.17	3.63	7.75	4.03	.82**
ANG										
Subtest	T1-T1 (same form)					T2-T2 (same form)				
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T1-T1}}$	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T2-T2}}$
VA	13.80	3.86	14.69	3.49	.74**	14.34	4.45	14.76	3.76	.82**
AR	11.05	4.28	11.98	4.24	.81**	10.78	4.66	11.15	4.37	.91**
WK	13.02	5.31	13.89	5.27	.88**	13.61	5.31	14.27	5.35	.94**
MK	9.61	4.77	11.02	4.63	.80**	10.90	4.17	12.22	3.60	.72**
RC	16.31	4.11	16.86	4.11	.77**	17.41	3.56	17.29	3.86	.86**
PS	8.30	3.53	8.46	3.25	.80**	9.34	3.60	9.59	3.56	.84**
TR	28.53	6.63	29.14	7.15	.72**	27.68	7.50	28.27	7.05	.77**
IC	13.45	6.70	14.71	7.08	.88**	13.41	7.05	14.02	6.62	.92**
BC	15.97	4.41	16.49	4.48	.69**	17.54	5.69	18.24	5.44	.90**
AI	10.17	5.12	11.08	5.16	.90**	9.87	4.49	10.10	4.73	.94**
AFRES										
Subtest	T1-T1 (same form)					T2-T2 (same form)				
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T1-T1}}$	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	$r_{\text{T2-T2}}$
VA	12.16	4.27	13.16	3.67	.71**	12.40	4.60	13.90	3.70	.84**
AR	10.32	3.58	11.47	3.78	.82**	9.15	3.15	9.50	2.80	.59**
WK	9.89	5.16	10.84	4.90	.72**	10.70	5.29	11.70	4.41	.93**
MK	10.53	3.75	11.68	3.84	.82**	9.15	3.60	10.85	3.62	.83**
RC	14.26	3.77	14.00	4.14	.87**	14.45	5.41	15.35	4.45	.93**
PS	7.74	2.51	7.79	2.46	.71**	8.00	3.03	7.75	2.97	.90**
TR	29.16	6.65	31.21	4.92	.44 [†]	25.65	7.28	26.50	8.02	.88**
IC	12.42	7.16	15.47	6.92	.73**	12.05	7.10	11.80	6.96	.96**
BC	14.37	6.72	16.05	6.50	.77**	15.05	6.00	16.20	5.74	.93**
AI	9.89	5.56	10.56	6.29	.94**	8.65	4.07	8.70	4.28	.86**

** $p \leq .010$; [†] $p \leq .050$.

Note. M_{Earlier} = Mean of the Earlier Test Administration; SD_{Earlier} = Standard Deviation of the Earlier Test Administration; M_{Later} = Mean of the Later Test Administration; SD_{Later} = Standard Deviation of the Later Test Administration; r = Correlation. NA means that the sample size was too small to generate stable statistics.

OTS-CIV: $N_{\text{T1-T1}} = 98$; $N_{\text{T2-T2}} = 73$;

OTS-AD: $N_{\text{T1-T1}} = 100$; $N_{\text{T2-T2}} = 83$;

AECP: $N_{\text{T1-T1}} = 0$; $N_{\text{T2-T2}} = 2$;

USAFA: $N_{\text{T1-T1}} = 82$; $N_{\text{T2-T2}} = 8$;

ROTC: $N_{\text{T1-T1}} = 351$; $N_{\text{T2-T2}} = 338$;

ANG: $N_{\text{T1-T1}} = 64$; $N_{\text{T2-T2}} = 41$;

AFRES: $N_{\text{T1-T1}} = 19$; $N_{\text{T2-T2}} = 20$.

For subtest-specific N , please refer to the Attachment 2.

APPENDIX O. Stratified Samples Between-Subj. Alt. Forms Reliability – All Items

OTS-CIV					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	15.17	4.56	15.04	4.68	.03 [†]
AR	13.51	5.08	13.30	5.27	.04*
WK	15.51	5.70	15.08	5.65	.07**
MK	14.02	5.25	14.11	5.11	.02
RC	17.03	4.08	17.69	4.39	.16**
PS	10.42	3.98	10.83	3.88	.10**
TR	27.88	6.45	27.41	6.63	.07**
IC	14.53	6.81	14.58	6.56	.01
BC	16.56	6.13	17.09	5.48	.09**
AI	9.22	4.43	8.62	4.01	.14**

Note. All except WK, PS, and TR use Welch's *t*-test.

OTS-AD					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	14.53	4.23	14.33	4.33	.05*
AR	12.49	4.66	12.03	4.79	.10**
WK	15.12	5.50	14.77	5.36	.07**
MK	11.67	4.52	11.82	4.40	.03
RC	17.15	3.84	17.25	4.01	.02
PS	8.79	3.44	9.47	3.36	.20**
TR	27.95	6.34	27.62	6.46	.05*
IC	14.04	6.70	14.01	6.40	.00
BC	15.80	6.02	16.72	5.42	.16**
AI	9.20	3.75	8.71	3.56	.13**

Note. All except AR, PS, and TR use Welch's *t*-test.

AECF					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	13.85	4.42	13.66	4.34	.04
AR	11.42	4.74	12.01	4.89	.12
WK	13.39	5.81	13.71	5.94	.05
MK	11.89	4.51	12.05	4.09	.04
RC	16.11	3.81	16.42	4.50	.08
PS	8.52	3.48	9.81	3.57	.37**
TR	26.98	7.01	26.45	6.75	.08
IC	13.62	6.46	14.19	6.26	.09
BC	14.73	6.45	15.80	5.45	.18
AI	7.87	3.42	7.87	3.40	.00

Note. RC and BC use Welch's *t*-test.

USAFA					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	15.83	3.82	16.03	3.96	.05*
AR	16.11	4.42	16.30	4.61	.04 [†]
WK	16.20	5.03	15.78	4.94	.08**
MK	18.46	4.09	18.10	3.90	.09**
RC	18.16	3.67	18.74	3.71	.16**
PS	13.24	3.21	13.72	3.12	.15**
TR	28.44	5.52	27.57	5.78	.15**
IC	15.81	6.30	15.48	6.20	.05*
BC	13.74	5.20	15.00	4.68	.25**
AI	10.78	3.87	10.20	3.49	.16**

Note. All except WK, RC, TR, and IC uses Welch's *t*-test.

ROTC					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	14.16	4.41	13.95	4.64	.05**
AR	12.91	5.01	12.86	5.17	.01
WK	13.72	5.54	13.49	5.37	.04**
MK	14.71	5.29	14.59	5.04	.02 [†]
RC	16.58	4.15	17.07	4.55	.11**
PS	10.91	3.92	11.26	3.92	.09**
TR	26.08	6.14	25.56	6.36	.08**
IC	14.40	6.81	14.36	6.52	.01
BC	14.31	5.70	15.27	5.08	.18**
AI	8.63	4.11	8.31	3.75	.08**

Note. All use Welch's *t*-test.

ANG					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	14.88	4.23	14.83	4.35	.01
AR	12.84	4.79	12.57	4.87	.06*
WK	15.36	5.54	15.19	5.38	.03
MK	12.35	4.82	12.53	4.53	.04 [†]
RC	17.48	3.85	17.84	3.93	.09**
PS	9.63	3.69	10.27	3.59	.18**
TR	28.50	6.38	28.18	6.64	.05*
IC	15.15	6.84	15.12	6.60	.00
BC	16.34	5.99	17.07	5.32	.13**
AI	10.86	4.85	10.32	4.57	.12**

Note. WK, MK, BC, and AI use Welch's *t*-test.

AFRES					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	14.52	4.38	14.32	4.42	.05
AR	12.46	4.79	12.33	5.11	.03
WK	15.08	5.78	15.07	5.59	.00
MK	12.27	4.98	12.57	4.80	.06
RC	17.12	4.02	17.39	4.05	.07 [†]
PS	9.33	3.80	9.97	3.73	.17**
TR	27.83	6.49	27.49	6.55	.05
IC	14.24	7.05	14.64	6.81	.06
BC	15.57	6.39	16.61	5.61	.17**
AI	10.67	5.07	9.89	4.57	.16**

Note. AR, BC, and AI use Welch's *t*-test.

[†] $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Overall Notes. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; ES = Effect Size = Cohen's d .

OTS-CIV: $N_{T1} = 7,352$; $N_{T2} = 7,284$;

OTS-AD: $N_{T1} = 4,967$; $N_{T2} = 4,843$;

AECP: $N_{T1} = 158$; $N_{T2} = 168$;

USAFA: $N_{T1} = 3,968$; $N_{T2} = 3,741$;

ROTC: $N_{T1} = 11,577$; $N_{T2} = 12,413$;

ANG: $N_{T1} = 3,585$; $N_{T2} = 3,288$;

AFRES: $N_{T1} = 1,345$; $N_{T2} = 1,215$.

For subtest-specific N , please refer to the Attachment 2.

APPENDIX P. Stratified Samples Between-Subj. Alt. Forms Reliability – Common Items

OTS-CIV					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	5.22	1.78	5.12	1.80	.05**
AR	3.84	1.96	3.88	1.97	.02
WK	5.08	2.40	5.12	2.43	.02
MK	4.74	2.21	4.71	2.15	.01
RC	NA	NA	NA	NA	NA
PS	3.75	1.75	3.83	1.76	.05**
TR	7.29	1.89	7.25	1.87	.02
IC	6.04	3.87	6.01	3.92	.01
BC	9.08	3.29	9.18	3.21	.03 [†]
AI	3.04	1.70	3.07	1.68	.02
OTS-AD					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	4.99	1.72	4.84	1.73	.09**
AR	3.47	1.79	3.42	1.84	.03
WK	4.87	2.39	4.85	2.37	.01
MK	3.84	1.90	3.78	1.89	.03
RC	NA	NA	NA	NA	NA
PS	3.12	1.62	3.22	1.65	.06**
TR	7.27	1.89	7.27	1.85	.00
IC	5.81	3.74	5.71	3.78	.03
BC	8.80	3.30	8.88	3.26	.02
AI	3.22	1.55	3.32	1.57	.06**
AECF					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	4.78	1.74	4.53	1.77	.14
AR	3.22	1.82	3.35	1.86	.07
WK	4.23	2.54	4.53	2.60	.11
MK	3.85	1.87	4.01	1.81	.08
RC	NA	NA	NA	NA	NA
PS	2.96	1.55	3.43	1.72	.29**
TR	6.99	1.98	6.98	2.04	.01
IC	5.46	3.62	5.79	3.70	.09
BC	8.17	3.53	8.62	3.28	.13
AI	2.81	1.46	2.91	1.46	.07
USAFA					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	5.44	1.63	5.31	1.62	.08**
AR	4.81	1.83	4.92	1.84	.06*
WK	5.46	2.15	5.36	2.21	.05*
MK	6.55	1.76	6.41	1.72	.08**
RC	NA	NA	NA	NA	NA
PS	5.00	1.46	5.11	1.41	.08**
TR	7.43	1.72	7.28	1.68	.09**
IC	6.77	3.68	6.56	3.77	.06*
BC	8.08	3.11	8.05	3.16	.01
AI	3.87	1.46	3.93	1.43	.04 [†]

ROTC					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	4.80	1.78	4.62	1.79	.10**
AR	3.76	1.91	3.80	1.94	.02 [†]
WK	4.51	2.31	4.58	2.29	.03*
MK	5.05	2.19	5.01	2.10	.02
RC	NA	NA	NA	NA	NA
PS	3.95	1.76	4.05	1.75	.06**
TR	6.78	1.87	6.72	1.87	.03*
IC	6.15	3.76	6.04	3.81	.03*
BC	8.18	3.28	8.24	3.26	.02
AI	3.02	1.59	3.08	1.58	.04**
ANG					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	5.06	1.74	4.94	1.75	.07**
AR	3.59	1.85	3.61	1.85	.01
WK	5.03	2.38	5.11	2.31	.03
MK	4.12	2.04	4.07	1.95	.02
RC	NA	NA	NA	NA	NA
PS	3.46	1.71	3.62	1.72	.09**
TR	7.42	1.87	7.40	1.87	.01
IC	6.45	3.89	6.43	3.96	.01
BC	9.04	3.24	9.07	3.21	.01
AI	3.64	1.79	3.70	1.79	.04
AFRES					
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	ES
VA	4.92	1.77	4.76	1.76	.09*
AR	3.46	1.84	3.59	1.94	.07 [†]
WK	4.94	2.43	5.07	2.40	.05
MK	4.04	2.09	4.15	2.00	.05
RC	NA	NA	NA	NA	NA
PS	3.40	1.72	3.48	1.75	.04
TR	7.25	1.90	7.24	1.85	.01
IC	5.98	3.98	6.09	4.02	.03
BC	8.64	3.48	8.92	3.36	.08*
AI	3.57	1.85	3.61	1.78	.02

[†] $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Notes. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; ES = Effect Size = Cohen's d ; NA = Not Available because RC had no common items.

OTS-CIV: $N_{T1} = 7,352$; $N_{T2} = 7,284$;

OTS-AD: $N_{T1} = 4,967$; $N_{T2} = 4,843$;

AECP: $N_{T1} = 158$; $N_{T2} = 168$;

USAFA: $N_{T1} = 3,968$; $N_{T2} = 3,741$;

ROTC: $N_{T1} = 11,577$; $N_{T2} = 12,413$;

ANG: $N_{T1} = 3,585$; $N_{T2} = 3,288$;

AFRES: $N_{T1} = 1,345$; $N_{T2} = 1,215$.

For subtest-specific N , please refer to the Attachment 2.

APPENDIX Q. Stratified Samples Within-Subj. Alt. Forms Reliability – All Items

OTS-CIV						
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	r_{T1-T2}	ES
VA	13.35	4.52	13.16	4.21	.66**	.05
AR	10.88	4.60	11.21	4.22	.67**	.09 [†]
WK	12.51	5.41	12.77	5.32	.78**	.07
MK	11.94	4.23	11.49	4.48	.65**	.12*
RC	15.63	4.29	15.15	4.14	.63**	.13**
PS	9.44	3.38	9.07	3.36	.65**	.13*
TR	26.61	6.88	26.66	6.94	.55**	.01
IC	14.38	6.39	14.10	6.81	.73**	.06
BC	16.10	5.30	15.53	5.50	.63**	.12*
AI	8.82	4.30	9.51	4.68	.72**	.20**
OTS-AD						
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	r_{T1-T2}	ES
VA	12.51	3.90	12.85	3.86	.52**	.09*
AR	10.04	3.93	10.28	3.75	.56**	.07 [†]
WK	12.48	4.94	12.86	5.14	.78**	.11**
MK	9.94	3.65	9.76	3.67	.47**	.05
RC	15.41	3.96	15.72	3.75	.59**	.09*
PS	8.12	2.90	7.28	2.96	.61**	.32**
TR	26.56	6.78	26.83	6.60	.56**	.04
IC	12.84	6.36	12.82	6.82	.72**	.01
BC	15.69	5.12	14.81	5.59	.67**	.20**
AI	7.89	3.30	8.72	3.76	.69**	.29**
AECF						
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	r_{T1-T2}	ES
VA	NA	NA	NA	NA	NA	NA
AR	NA	NA	NA	NA	NA	NA
WK	NA	NA	NA	NA	NA	NA
MK	NA	NA	NA	NA	NA	NA
RC	NA	NA	NA	NA	NA	NA
PS	NA	NA	NA	NA	NA	NA
TR	NA	NA	NA	NA	NA	NA
IC	NA	NA	NA	NA	NA	NA
BC	NA	NA	NA	NA	NA	NA
AI	NA	NA	NA	NA	NA	NA
USAFA						
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	r_{T1-T2}	ES
VA	14.81	4.08	14.79	3.72	.50**	.00
AR	13.62	4.27	13.86	4.06	.58**	.06
WK	14.07	4.81	14.71	4.94	.75**	.19**
MK	15.61	3.78	16.05	3.81	.52**	.12*
RC	17.61	4.13	17.35	3.76	.48**	.07
PS	12.02	2.92	11.61	2.92	.56**	.15**
TR	25.05	6.49	26.66	5.59	.16**	.21**
IC	12.64	5.69	13.14	6.00	.18**	.07
BC	13.17	4.34	12.48	4.57	.37**	.14**
AI	8.98	2.93	10.10	3.42	.19**	.28**

ROTC						
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	r_{T1-T2}	ES
VA	11.66	4.39	11.91	4.12	.59**	.07**
AR	10.17	4.49	10.21	4.24	.70**	.01
WK	10.49	4.84	10.63	5.05	.76**	.04*
MK	11.92	4.47	11.90	4.64	.70**	.01
RC	14.47	4.56	14.34	4.11	.62**	.04 [†]
PS	9.26	3.45	8.88	3.43	.69**	.14**
TR	24.51	6.48	25.09	6.20	.52**	.09**
IC	12.81	6.36	12.66	6.76	.71**	.03
BC	14.23	4.98	13.11	5.47	.62**	.25**
AI	7.53	3.54	7.91	4.05	.65**	.12**
ANG						
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	r_{T1-T2}	ES
VA	12.98	4.00	12.91	3.66	.50**	.02
AR	10.27	4.11	10.50	3.94	.56**	.06
WK	12.51	5.02	12.24	5.19	.78**	.08
MK	10.19	4.06	10.01	3.87	.55**	.05
RC	15.90	3.98	15.72	3.74	.51**	.05
PS	8.84	3.03	8.32	3.13	.63**	.20**
TR	28.02	6.54	28.44	6.67	.49**	.06
IC	14.93	6.61	14.98	6.85	.76**	.01
BC	16.72	5.23	15.55	5.93	.64**	.24**
AI	10.44	4.77	11.10	5.08	.78**	.20**
AFRES						
Subtest	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	r_{T1-T2}	ES
VA	13.15	4.01	12.78	4.32	.57**	.10
AR	10.04	4.07	10.06	4.14	.57**	.01
WK	12.78	5.35	11.97	5.49	.77**	.22**
MK	10.88	4.08	10.41	3.78	.59**	.13 [†]
RC	15.24	4.23	15.47	3.80	.61**	.06
PS	8.24	3.15	7.67	3.10	.64**	.22**
TR	27.18	6.78	27.06	6.26	.58**	.02
IC	14.28	6.89	13.75	7.07	.70**	.10
BC	15.75	5.48	14.67	6.01	.69**	.24**
AI	9.77	4.94	10.15	4.90	.76**	.11

[†] $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; r = Correlation; T2; ES = Effect Size = Cohen's d .

OTS-CIV: $N_{T1} = 395$; $N_{T2} = 395$;

OTS-AD: $N_{T1} = 613$; $N_{T2} = 613$;

AECP: $N_{T1} = 6$; $N_{T2} = 6$;

USAFA: $N_{T1} = 411$; $N_{T2} = 411$;

ROTC: $N_{T1} = 3,010$; $N_{T2} = 3,010$;

ANG: $N_{T1} = 391$; $N_{T2} = 391$;

AFRES: $N_{T1} = 156$; $N_{T2} = 156$.

For subtest-specific N , please refer to the Attachment 2.

APPENDIX R. Stratified Samples Retesting Effects on the Same Form – T1T1

OTS-CIV							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	13.39	4.73	14.09	4.68	-2.91	97	.29**
AR	12.41	5.05	13.17	4.82	-2.96	97	.30**
WK	12.84	5.91	13.79	6.00	-3.59	97	.36**
MK	12.61	5.17	13.79	5.16	-4.07	97	.41**
RC	15.41	4.44	16.28	4.16	-3.09	97	.31**
PS	9.69	3.82	9.97	3.86	-1.40	96	.14
TR	26.98	6.73	28.23	6.67	-3.14	96	.32**
IC	13.79	7.15	15.44	7.08	-3.74	96	.38**
BC	15.65	6.43	16.91	6.46	-3.25	96	.33**
AI	9.18	4.76	9.80	4.85	-3.27	96	.33**
OTS-AD							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	13.58	3.71	14.43	3.59	-3.87	99	.39**
AR	11.17	4.06	11.95	4.30	-2.84	99	.28**
WK	13.05	5.55	13.95	5.37	-3.70	99	.37**
MK	10.37	3.86	11.78	3.56	-4.89	99	.49**
RC	16.11	4.23	17.06	3.81	-3.89	99	.39**
PS	8.36	3.21	8.41	3.46	-.77	98	.08
TR	28.09	6.34	29.08	6.56	-2.43	98	.24*
IC	14.10	6.15	15.35	6.32	-4.03	98	.40**
BC	15.66	5.91	16.63	6.10	-2.52	98	.25*
AI	8.77	3.90	9.47	4.22	-3.75	98	.38**
AECP							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	NA	NA	NA	NA	NA	NA	NA
AR	NA	NA	NA	NA	NA	NA	NA
WK	NA	NA	NA	NA	NA	NA	NA
MK	NA	NA	NA	NA	NA	NA	NA
RC	NA	NA	NA	NA	NA	NA	NA
PS	NA	NA	NA	NA	NA	NA	NA
TR	NA	NA	NA	NA	NA	NA	NA
IC	NA	NA	NA	NA	NA	NA	NA
BC	NA	NA	NA	NA	NA	NA	NA
AI	NA	NA	NA	NA	NA	NA	NA
USAFA							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	15.11	3.46	15.48	3.51	-2.14	81	.24*
AR	15.21	4.05	15.23	4.13	-.15	81	.02
WK	14.38	4.25	14.38	4.52	.10	80	.01
MK	17.38	4.03	17.56	3.93	-1.62	80	.18
RC	16.91	4.32	17.01	4.17	-.73	80	.08
PS	12.74	3.06	12.95	2.95	-1.77	79	.20†
TR	26.78	5.46	27.04	5.62	-1.45	79	.16
IC	13.79	6.78	14.55	6.77	-2.40	79	.27*
BC	12.40	5.12	12.77	5.14	-1.84	79	.21†
AI	10.06	3.62	10.59	3.60	-2.86	79	.32**

ROTC							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	10.59	3.77	12.72	3.65	-13.91	350	.74**
AR	9.51	3.89	10.50	4.03	-6.24	350	.33**
WK	9.51	4.44	10.89	4.64	-8.47	350	.45**
MK	11.26	4.33	12.40	4.15	-6.90	350	.37**
RC	13.62	3.85	15.18	3.87	-9.80	350	.52**
PS	8.49	3.10	8.70	3.16	-1.69	349	.09†
TR	24.63	6.13	26.67	6.69	-7.33	350	.39**
IC	11.40	6.83	13.95	7.16	-10.39	348	.56**
BC	12.44	5.45	14.47	5.89	-9.81	350	.52**
AI	7.41	4.15	8.38	4.45	-7.12	349	.38**
ANG							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	13.80	3.86	14.69	3.49	-2.69	63	.34**
AR	11.05	4.28	11.98	4.24	-2.83	63	.35**
WK	13.02	5.31	13.89	5.27	-2.67	63	.33**
MK	9.61	4.77	11.02	4.63	-3.51	62	.44**
RC	16.31	4.11	16.86	4.11	-1.56	63	.19
PS	8.30	3.53	8.46	3.25	-.40	62	.05
TR	28.53	6.63	29.14	7.15	-.66	62	.08
IC	13.45	6.70	14.71	7.08	-2.50	62	.31*
BC	15.97	4.41	16.49	4.48	-.84	62	.11
AI	10.17	5.12	11.08	5.16	-2.81	61	.36**
AFRES							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	12.16	4.27	13.16	3.67	-1.42	18	.33
AR	10.32	3.58	11.47	3.78	-2.25	18	.52*
WK	9.89	5.16	10.84	4.90	-1.09	18	.25
MK	10.53	3.75	11.68	3.84	-2.20	18	.51*
RC	14.26	3.77	14.00	4.14	.55	18	.13
PS	7.74	2.51	7.79	2.46	-.12	18	.03
TR	29.16	6.65	31.21	4.92	-1.42	18	.32
IC	12.42	7.16	15.47	6.92	-2.57	18	.59*
BC	14.37	6.72	16.05	6.50	-1.63	18	.37
AI	9.89	5.56	10.56	6.29	-1.27	17	.30

† $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M_{Earlier} = Mean of Earlier Test Administration; SD_{Earlier} = Standard Deviation of Earlier Test Administration; M_{Later} = Mean of Later Test Administration; SD_{Later} = Standard Deviation of Later Test Administration; $t = t$ -test; df = Degrees of Freedom; ES = Effect Size = Cohen's d . NA means that the sample size was too small to generate stable statistics.

OTS-CIV: $N = 98$;

OTS-AD: $N = 100$;

AECP: $N = 0$;

USAFA: $N = 82$;

ROTC: $N = 351$;

ANG: $N = 64$;

AFRES: $N = 19$.

For subtest-specific N , please refer to the Attachment 2.

APPENDIX S. Stratified Samples Retesting Effects on the Same Form – T2T2

OTS-CIV							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	14.71	4.52	14.81	4.42	-.53	72	.06
AR	13.36	5.01	13.47	5.06	-.41	72	.05
WK	14.63	5.76	14.81	5.80	-.71	72	.08
MK	13.19	5.19	13.73	4.99	-2.19	72	.26*
RC	17.26	4.07	17.41	3.92	-1.05	72	.12
PS	10.51	3.09	10.55	3.12	-.31	72	.04
TR	26.44	6.45	26.88	7.28	-1.18	72	.14
IC	14.73	6.60	14.86	6.83	-.48	72	.06
BC	16.86	5.35	17.37	5.25	-1.88	72	.22†
AI	8.16	4.21	8.45	4.39	-1.59	72	.19
OTS-AD							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	12.34	4.56	13.41	4.08	-3.25	82	.36**
AR	10.08	4.29	11.11	4.40	-3.81	82	.42**
WK	13.23	5.58	13.65	5.47	-1.78	82	.20†
MK	10.53	3.71	11.33	3.65	-2.63	82	.29**
RC	15.08	4.69	15.66	4.56	-2.43	82	.27*
PS	8.45	2.60	8.37	2.85	.36	82	.04
TR	26.27	6.96	27.80	6.36	-2.63	82	.29**
IC	11.28	5.61	12.30	6.15	-2.24	82	.25*
BC	14.13	4.87	15.13	4.94	-2.99	82	.33**
AI	7.64	3.09	8.18	3.39	-2.32	82	.25*
AECP							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	NA	NA	NA	NA	NA	NA	NA
AR	NA	NA	NA	NA	NA	NA	NA
WK	NA	NA	NA	NA	NA	NA	NA
MK	NA	NA	NA	NA	NA	NA	NA
RC	NA	NA	NA	NA	NA	NA	NA
PS	NA	NA	NA	NA	NA	NA	NA
TR	NA	NA	NA	NA	NA	NA	NA
IC	NA	NA	NA	NA	NA	NA	NA
BC	NA	NA	NA	NA	NA	NA	NA
AI	NA	NA	NA	NA	NA	NA	NA
USAFA							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	NA	NA	NA	NA	NA	NA	NA
AR	NA	NA	NA	NA	NA	NA	NA
WK	NA	NA	NA	NA	NA	NA	NA
MK	NA	NA	NA	NA	NA	NA	NA
RC	NA	NA	NA	NA	NA	NA	NA
PS	NA	NA	NA	NA	NA	NA	NA
TR	NA	NA	NA	NA	NA	NA	NA
IC	NA	NA	NA	NA	NA	NA	NA
BC	NA	NA	NA	NA	NA	NA	NA
AI	NA	NA	NA	NA	NA	NA	NA

ROTC							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	10.37	4.08	12.79	4.01	-14.07	337	.77**
AR	9.06	4.03	10.09	4.03	-6.35	337	.35**
WK	9.83	4.74	11.05	4.73	-7.38	337	.40**
MK	10.66	4.13	11.90	3.76	-7.47	337	.41**
RC	13.57	4.55	15.22	4.46	-8.75	337	.48**
PS	8.82	3.44	9.11	3.42	-2.43	336	.13*
TR	23.32	6.09	24.96	7.21	-4.97	334	.27**
IC	11.68	6.15	13.37	6.54	-8.03	336	.44**
BC	13.22	4.56	15.17	4.94	-10.57	335	.58**
AI	7.17	3.63	7.75	4.03	-4.29	335	.23**
ANG							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	14.34	4.45	14.76	3.76	-1.05	40	.16
AR	10.78	4.66	11.15	4.37	-1.19	40	.19
WK	13.61	5.31	14.27	5.35	-2.35	40	.37*
MK	10.90	4.17	12.22	3.60	-2.86	40	.45**
RC	17.42	3.56	17.29	3.86	.39	40	.06
PS	9.34	3.60	9.59	3.56	-.77	40	.12
TR	27.68	7.50	28.27	7.05	-.76	40	.12
IC	13.42	7.05	14.02	6.62	-1.40	40	.22
BC	17.54	5.69	18.24	5.44	-1.77	40	.28†
AI	9.87	4.49	10.10	4.73	-.87	38	.14
AFRES							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	12.40	4.60	13.90	3.70	-2.66	19	.59*
AR	9.15	3.15	9.50	2.80	-.58	19	.13
WK	10.70	5.29	11.70	4.41	-2.24	19	.50*
MK	9.15	3.60	10.85	3.62	-3.61	19	.81**
RC	14.45	5.41	15.35	4.45	-1.96	19	.44†
PS	8.00	3.03	7.75	2.97	.84	19	.19
TR	25.65	7.28	26.50	8.02	-1.02	19	.23
IC	12.05	7.10	11.80	6.96	.54	19	.12
BC	15.05	6.00	16.20	5.74	-2.26	19	.50*
AI	8.65	4.07	8.70	4.28	-.10	19	.02

† $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M_{Earlier} = Mean of Earlier Test Administration; SD_{Earlier} = Standard Deviation of Earlier Test Administration; M_{Later} = Mean of Later Test Administration; SD_{Later} = Standard Deviation of Later Test Administration; $t = t$ -test; df = Degrees of Freedom; ES = Effect Size = Cohen's d . NA means that the sample size was too small to generate stable statistics.

OTS-CIV: $N = 73$;

OTS-AD: $N = 83$;

AECP: $N = 2$;

USAFA: $N = 8$;

ROTC: $N = 338$;

ANG: $N = 41$;

AFRES: $N = 20$.

For subtest-specific N , please refer to the Attachment 2.

APPENDIX T. Stratified Sample Retesting Effects on the Parallel Forms

OTS-CIV							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	12.47	4.43	14.04	4.16	-9.51	394	.48**
AR	10.43	4.41	11.66	4.34	-7.13	394	.36**
WK	12.03	5.29	13.25	5.37	-7.26	394	.37**
MK	11.07	4.38	12.35	4.24	-7.37	394	.37**
RC	15.18	4.26	15.60	4.17	-2.30	394	.12*
PS	9.15	3.34	9.36	3.40	-1.41	393	.07
TR	25.19	6.55	28.07	6.96	-9.65	393	.49**
IC	13.18	6.35	15.31	6.68	-9.66	393	.49**
BC	14.82	5.34	16.80	5.29	-9.30	393	.47**
AI	8.63	4.14	9.71	4.79	-6.49	393	.33**
OTS-AD							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	11.98	3.86	13.38	3.79	-9.78	612	.40**
AR	9.54	3.71	10.77	3.88	-8.93	612	.36**
WK	12.07	5.04	13.27	4.98	-9.47	612	.38**
MK	9.10	3.58	10.60	3.58	-10.68	612	.43**
RC	15.22	3.93	15.91	3.76	-4.96	612	.20**
PS	7.70	2.84	7.70	3.08	.06	611	.00
TR	25.82	6.60	27.57	6.67	-7.25	610	.29**
IC	11.98	6.38	13.68	6.69	-8.96	609	.36**
BC	14.49	5.30	16.01	5.35	-8.93	610	.36**
AI	8.01	3.44	8.60	3.65	-5.01	607	.20**
AECF							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	NA	NA	NA	NA	NA	NA	NA
AR	NA	NA	NA	NA	NA	NA	NA
WK	NA	NA	NA	NA	NA	NA	NA
MK	NA	NA	NA	NA	NA	NA	NA
RC	NA	NA	NA	NA	NA	NA	NA
PS	NA	NA	NA	NA	NA	NA	NA
TR	NA	NA	NA	NA	NA	NA	NA
IC	NA	NA	NA	NA	NA	NA	NA
BC	NA	NA	NA	NA	NA	NA	NA
AI	NA	NA	NA	NA	NA	NA	NA
USAFA							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	14.10	3.99	15.49	3.68	-7.73	410	.38**
AR	13.13	4.06	14.36	4.18	-6.81	410	.34**
WK	13.70	4.91	15.08	4.77	-8.61	406	.43**
MK	15.27	3.80	16.38	3.72	-6.24	407	.31**
RC	16.64	4.16	18.32	3.54	-9.27	405	.46**
PS	11.94	2.98	11.70	2.87	1.68	400	.08†
TR	23.65	5.48	28.07	5.90	-13.23	400	.66**
IC	10.20	5.22	15.59	5.17	-20.79	400	1.04**
BC	11.40	4.03	14.26	4.44	-13.78	400	.69**
AI	8.18	2.84	10.90	3.02	-17.03	400	.85**

ROTC							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	10.93	4.11	12.64	4.24	-26.89	3008	.49**
AR	9.74	4.35	10.64	4.34	-15.15	3008	.28**
WK	10.06	4.84	11.06	5.00	-16.68	3006	.30**
MK	11.38	4.55	12.45	4.50	-17.33	3007	.32**
RC	13.83	4.34	14.97	4.26	-17.23	3006	.31**
PS	8.98	3.42	9.15	3.46	-3.35	3001	.06**
TR	23.71	5.91	25.89	6.58	-20.49	3000	.37**
IC	11.63	6.19	13.84	6.74	-26.72	2995	.49**
BC	12.66	5.02	14.67	5.30	-26.04	2998	.48**
AI	7.23	3.48	8.21	4.05	-17.33	2991	.32**
ANG							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	12.21	3.69	13.68	3.83	-8.24	390	.42**
AR	9.71	3.85	11.06	4.09	-7.54	390	.38**
WK	11.69	5.05	13.05	5.07	-8.68	390	.44**
MK	9.27	3.79	10.93	3.96	-9.72	390	.49**
RC	15.40	4.00	16.22	3.68	-4.34	390	.22**
PS	8.51	3.14	8.64	3.04	-.98	389	.05
TR	26.96	6.27	29.50	6.70	-8.17	390	.41**
IC	14.28	6.49	15.63	6.89	-5.95	389	.30**
BC	15.16	5.40	17.12	5.67	-8.54	390	.43**
AI	10.31	4.60	11.24	5.21	-5.66	388	.29**
AFRES							
Subtest	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	t	df	ES
VA	12.29	4.08	13.63	4.16	-4.58	155	.37**
AR	9.44	3.94	10.65	4.17	-4.22	155	.34**
WK	11.57	5.51	13.18	5.24	-5.84	155	.47**
MK	9.94	3.92	11.36	3.83	-5.38	155	.43**
RC	15.33	4.01	15.38	4.03	-.18	155	.01
PS	7.74	3.02	8.18	3.24	-2.06	155	.17*
TR	26.08	6.31	28.17	6.58	-4.65	155	.37**
IC	12.83	6.71	15.21	7.04	-6.07	154	.49**
BC	14.15	5.68	16.27	5.69	-6.32	155	.51**
AI	9.50	4.63	10.42	5.15	-3.43	155	.27**

[†] $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M_{Earlier} = Mean of the Earlier Test Administration; SD_{Earlier} = Standard Deviation of Earlier Administration; M_{Later} = Mean of the Later Test Administration; SD_{Later} = Standard Deviation of Later Administration; t = t -test; df = Degrees of Freedom; ES = Effect Size = Cohen's d .

OTS-CIV: $N_{\text{Earlier}} = 395$; $N_{\text{Later}} = 395$;

OTS-AD: $N_{\text{Earlier}} = 613$; $N_{\text{Later}} = 613$;

AECP: $N_{\text{Earlier}} = 6$; $N_{\text{Later}} = 6$;

USAFA: $N_{\text{Earlier}} = 411$; $N_{\text{Later}} = 411$;

ROTC: $N_{\text{Earlier}} = 3,010$; $N_{\text{Later}} = 3,010$;

ANG: $N_{\text{Earlier}} = 391$; $N_{\text{Later}} = 391$;

AFRES: $N_{\text{Earlier}} = 156$; $N_{\text{Later}} = 156$.

For subtest-specific N , please refer to the Attachment 2.

APPENDIX U. Stratified Samples Stability Analysis

OTS-CIV												
Subtest	T1						T2					
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.
VA	15.14	4.63	14.76	4.60	.08*	.38	14.86	4.77	14.90	4.64	.01	.03
AR	13.53	5.11	13.33	4.98	.04	.21	13.24	5.37	13.18	5.27	.01	.05
WK	15.34	5.77	15.29	5.70	.01	.05	14.96	5.76	14.85	5.63	.02	.11
MK	13.94	5.21	14.12	5.23	.04	.18	13.87	5.25	14.23	5.17	.07†	.36
RC	16.78	4.13	16.99	4.06	.05	.21	17.49	4.55	17.65	4.47	.03	.16
PS	10.15	4.01	10.61	4.02	.11**	.46	10.66	3.90	10.91	3.96	.06	.25
TR	27.54	6.52	27.78	6.51	.04	.24	27.32	6.69	27.37	6.46	.01	.05
IC	14.29	6.90	14.32	6.83	.01	.04	14.24	6.38	14.61	6.65	.06	.36
BC	16.29	6.10	16.42	6.23	.02	.13	16.93	5.39	17.06	5.48	.02	.13
AI	8.74	4.19	9.48	4.48	.17**	.74	8.04	3.62	9.07	4.07	.27**	1.03
OTS-AD												
Subtest	T1						T2					
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.
VA	14.72	4.19	14.52	4.20	.05	.20	14.33	4.35	14.42	4.32	.02	.09
AR	12.52	4.71	12.41	4.64	.03	.12	12.27	4.85	12.12	4.68	.03	.16
WK	15.24	5.51	15.24	5.51	.00	.01	14.64	5.49	15.10	5.28	.09*	.46
MK	11.53	4.60	11.69	4.49	.03	.16	11.84	4.58	12.12	4.28	.06	.28
RC	17.16	3.83	17.18	3.86	.00	.02	17.24	3.97	17.33	4.04	.02	.09
PS	8.68	3.47	8.98	3.54	.08*	.30	9.47	3.43	9.53	3.41	.02	.06
TR	28.45	6.37	27.61	6.33	.13**	.83	28.33	6.34	27.47	6.21	.14**	.86
IC	14.24	6.75	14.00	6.68	.04	.24	14.01	6.41	14.25	6.41	.04	.24
BC	16.33	6.03	15.60	6.03	.12**	.74	16.98	5.38	16.76	5.44	.04	.21
AI	9.29	3.79	9.36	3.81	.02	.07	8.65	3.60	8.78	3.51	.04	.13
AECF												
Subtest	T1						T2					
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.
VA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
WK	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
MK	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RC	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
TR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
IC	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
BC	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AI	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

USAFA												
Subtest	T1						T2					
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.
VA	16.48	3.65	15.36	3.80	.30**	1.12	16.62	3.68	15.45	4.11	.30**	1.18
AR	16.64	4.26	15.68	4.51	.22**	.96	16.70	4.51	15.71	4.68	.21**	.98
WK	17.15	4.94	15.44	5.00	.34**	1.71	16.42	4.79	15.13	4.98	.26**	1.29
MK	18.71	4.03	18.26	4.02	.11*	.45	18.07	3.88	18.10	3.88	.01	.03
RC	18.75	3.34	17.60	3.90	.32**	1.15	19.12	3.37	18.30	3.98	.22**	.82
PS	13.66	3.03	12.94	3.31	.22**	.71	13.84	3.16	13.63	3.10	.07	.21
TR	29.12	5.61	28.15	5.50	.17**	.96	28.16	5.76	26.96	5.50	.21**	1.20
IC	16.76	5.95	15.42	6.41	.22**	1.33	15.95	6.06	14.69	6.21	.21**	1.26
BC	13.67	5.37	13.40	5.20	.05	.26	14.86	4.77	14.64	4.45	.05	.21
AI	11.52	3.80	9.82	3.85	.44**	1.70	10.81	3.55	9.35	3.37	.42**	1.47
ROTC												
Subtest	T1						T2					
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.
VA	14.21	4.53	14.05	4.38	.04	.16	13.72	4.65	13.94	4.53	.05	.50
AR	12.88	5.07	12.74	4.94	.03	.14	12.62	5.19	12.96	5.10	.07*	.41
WK	13.61	5.67	13.52	5.49	.02	.09	13.17	5.50	13.55	5.24	.07*	.64
MK	14.75	5.31	14.62	5.30	.03	.13	14.45	5.12	14.67	4.99	.05	.11
RC	16.48	4.29	16.58	4.05	.03	.11	16.75	4.76	17.21	4.41	.10**	.12
PS	10.67	4.00	10.97	3.88	.07*	.29	11.02	3.91	11.44	3.92	.11**	.28
TR	26.18	6.22	25.79	6.07	.06†	.38	25.87	6.37	25.53	6.28	.05	.72
IC	14.61	6.89	14.26	6.81	.05	.35	14.36	6.54	14.43	6.49	.01	.36
BC	14.30	5.62	14.37	5.65	.01	.07	15.21	5.05	15.33	5.05	.02	.47
AI	8.68	4.07	8.70	4.12	.00	.02	8.25	3.75	8.35	3.77	.03	.09
ANG												
Subtest	T1						T2					
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.
VA	14.76	4.18	14.74	4.22	.01	.03	14.63	4.39	14.78	4.35	.03	.14
AR	12.85	4.71	12.75	4.91	.02	.10	12.45	4.68	12.39	4.91	.01	.06
WK	15.24	5.59	15.12	5.49	.02	.11	14.67	5.46	15.36	5.34	.13*	.69
MK	12.11	4.70	12.48	5.00	.08	.37	12.24	4.44	12.50	4.68	.06	.26
RC	17.19	3.89	17.56	3.77	.10*	.37	17.57	4.10	17.93	3.85	.09†	.36
PS	9.44	3.73	9.66	3.67	.06	.22	10.20	3.63	10.32	3.66	.03	.12
TR	28.66	6.21	28.28	6.54	.06	.38	28.36	6.64	27.77	6.63	.09†	.59
IC	15.23	6.82	14.63	6.99	.09†	.60	15.08	6.69	14.91	6.55	.02	.16
BC	16.33	5.94	16.05	6.17	.05	.28	17.04	5.31	16.85	5.43	.03	.19
AI	10.80	4.93	10.82	4.88	.00	.02	10.32	4.61	10.30	4.45	.00	.02

AFRES												
Subtest	T1						T2					
	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.	M_{Earlier}	SD_{Earlier}	M_{Later}	SD_{Later}	ES	Abs. M Diff.
VA	14.77	4.44	14.53	4.41	.05	.23	13.84	4.60	14.76	4.43	.20*	.92
AR	12.23	4.75	12.60	4.82	.08	.38	12.05	5.19	12.79	5.09	.14	.74
WK	15.29	5.83	15.28	5.69	.00	.01	14.26	5.61	15.71	5.55	.26**	1.45
MK	12.09	4.94	12.57	5.04	.10	.48	12.30	4.86	12.97	4.96	.14	.68
RC	16.93	4.17	17.15	4.09	.05	.22	17.26	4.14	17.80	3.93	.13	.54
PS	9.13	3.78	9.57	3.91	.11	.43	9.75	3.62	9.98	3.79	.06	.24
TR	27.65	6.57	28.04	6.51	.06	.39	27.83	6.83	27.55	6.38	.04	.28
IC	14.06	7.12	14.36	7.08	.04	.29	14.18	6.96	14.83	6.74	.10	.65
BC	15.19	6.62	15.42	6.60	.03	.23	16.90	5.38	16.78	6.04	.02	.11
AI	10.40	5.02	10.82	5.07	.08	.41	9.28	4.42	10.23	4.68	.21*	.94

[†] $p \leq .100$; * $p \leq .050$; ** $p \leq .010$; unmarked are non-significant p .

Note. M_{Earlier} = Mean of Earlier Group; SD_{Earlier} = Standard Deviation of Earlier Group; M_{Later} = Mean of Later Group; SD_{Later} = Standard Deviation of Later Group; ES = Effect Size = Cohen's d ; Abs. M Diff. = Absolute Mean Difference. NA means that the sample size was too small to generate stable statistics.

Total T1: OTS-CIV $N = 1,500$; OTS-AD $N = 1,300$; AECP $N_{\text{Earlier}} = 54$, $N_{\text{Later}} = 26$; USAFA $N = 1,000$; ROTC $N = 1,900$; ANG $N = 900$; AFRES $N = 320$.

Total T2: OTS-CIV $N = 1,200$; OTS-AD $N = 1,200$; AECP $N_{\text{Earlier}} = 34$, $N_{\text{Later}} = 45$; USAFA $N = 800$; ROTC $N = 1,800$; ANG $N = 800$; AFRES $N = 260$.

For subtest-specific N , please refer to the Attachment 2.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

ABM	Air Battle Manager
Abs. M Diff.	Absolute Mean Difference
AD-Other	Active Duty Other Accession Source
AECP	Airman Education and Commissioning Program
AFOQT	Air Force Officer Qualifying Test
AFPC/DSYX	Air Force Personnel Center Strategic Research and Assessments Branch
AFPTP	Air Force Personnel Testing Program
AFRES	Air Force Reserve
AFROTC	Air Force Reserve Officer Training Corps
AGFI	Adjusted Goodness of Fit Index
AI	Aviation Information
AIAN	American Indian/Alaska Native
AIC	Akaike Information Criterion
ANG	Air National Guard
AR	Arithmetic Reasoning
BC	Block Counting
BIC	Bayesian Information Criterion
CFA	Confirmatory Factor Analysis
CFI	Confirmatory Factor Index
Chi Square (χ^2)	Chi Square
Chi Square Diff. (χ^2 diff.)	Chi Square Difference between the Models
Confidence Intervals (CI)	
CIV-Other	Civilian Other Accession Source
Cohen's d (d)	Effect Size
Cronbach's alpha	Cronbach's alpha
CSO	Combat Systems Officer
CTT	Classical Test Theory
Delta R^2 (ΔR^2)	Change in R^2 when a variable is added or removed from a regression
df	degrees of freedom
df diff.	Degrees of Freedom Difference between the Models
g	General intelligence factor
GFI	Goodness of Fit Index
GLS	Generalized Least Squares Estimation
GS	General Science
HF	Hidden figures

IC	Instrument Comprehension
IER	Insufficient Effort Responding
IRT	Item Response Theory
IST	Infoscitex
ITC	Item-Total Correlation
LMEM	Linear Mixed Effects Model
<i>M</i>	Mean
MAR	Missing at random
Max	Maximum
MCAR	Missing completely at random
MI	Measurement Invariance
Min	Minimum
MK	Math Knowledge
ML	Maximum Likelihood
MNAR	Missing not at random
<i>N</i>	Sample Size
NA	Not Available
NHPI	Native Hawaiian/Other Pacific Islander
ns	Non-significant
Omega ²	Effect Size
OTS	Officer Training School
OTS-AD	Officer Training School-Active Duty
OTS-CIV	Officer Training School-Civilian
<i>p</i>	Probability value
PII	Personally Identifiable Information
PS	Physical Science
<i>p</i> -value	Proportion of correctly endorsed items in the test
PWS	Performance Work Statement
<i>r</i>	Correlation
<i>R</i> ²	R-Squared (Proportion of Variance Explained)
RB	Rotated Blocks
RC	Reading Comprehension
RGL	Reading Grade Level
RMSEA	Root Mean Square Error of Approximation
ROTC	Reserve Officer Training Corps
RPA	Remotely-Piloted Aircraft
RQ	Research Question
SD	Standard Deviation
SDI-O	Self-Description Inventory - Officers
<i>Sig.</i>	Significance Level

SJT	Situational Judgment Test
SPARKID	A dummy ID number used to assist in identifying data without the use of PII
SRMR	Standardized Root Mean Square Residual
t	A symbol used to denote the t-statistic
T1	Air Force Officer Qualifying Test Form T Form 1
T2	Air Force Officer Qualifying Test Form T Form 2
TCO-TE	Test Control Officer-Test Evaluator
TIF	Test Information Function
TLI	Tucker Lewis Index
TR	Table Reading
UPT	Undergraduate Pilot Training
URT	Undergraduate RPA Training
USAF	United States Air Force
USAFA	United States Air Force Academy
VA	Verbal Analogies
WK	Word Knowledge
WLS	Weighted Least Squares Estimation