



AFRL-RH-WP-TR-2022-0037

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

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

**Thomas R. Carretta
Performance Optimization Branch
Wright-Patterson AFB, OH**

**June 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-2022-0037 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

CARRETTA.THOMAS.R.1228984929
AS.R.1228984929
Digitally signed by
CARRETTA.THOMAS.R.1228984929
29
Date: 2022.07.12 08:42:01 -04'00'

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

MCKINLEY.RICHARD.A.1262239710
RD.A.1262239710
Digitally signed by
MCKINLEY.RICHARD.A.1262239710
0
Date: 2022.07.18 15:50:48 -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) 03-06-22		2. REPORT TYPE Interim		3. DATES COVERED (From - To) 1 Nov 2020 – 13 Jan 2021	
4. TITLE AND SUBTITLE Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: Item-Level Analyses				5a. CONTRACT NUMBER FA8650-20-F-4094	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER 00000F	
6. AUTHOR(S) Julia L. Walsh ^a , Michael F. Brady ^a , Montana R. Woolley ^a , and 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-2022-0037	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution Statement A: Approved for public release.					
13. SUPPLEMENTARY NOTES Report contains color. AFRL-2022-3143, Cleared 5 July 2022					
14. ABSTRACT This report describes item-level psychometric results of the 10 cognitive subtests on the Air Force Officer Qualifying Test (AFOQT) Form T, versions T1 and T2. The following were analyzed for each subtest: item difficulty, item discriminability, internal consistency if item is deleted, omission rates, overt guessing, distractor analysis, common items analysis, mean score subgroup differences, and item drift. Results show no major issues with the cognitive subtests. We found no practical differences between versions T1 and T2, therefore no equating was necessary. Overall, the mean score subgroup differences associated with items varied from small to large. Form T seemed to have better psychometric properties than were reported for Form S. These item-level results provide foundation for subtest-level analyses and the essential information which will allow the subtests to be improved in future AFOQT versions.					
15. KEY WORDS Air Force Officer Qualifying Test, AFOQT, item-level analyses					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT: SAR	18. NUMBER OF PAGES 139	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

1.0	INTRODUCTION	1
1.1	Brief History of the AFOQT	1
1.2	AFOQT Form T	1
1.3	Previous and Current Psychometric Evaluations	2
1.3.1	Morath et al. (2008)	2
1.3.2	Carretta et al. (2016)	2
1.3.3	Aguilar (2017).....	3
1.3.4	Note on Form S and Form T	3
1.3.5	The Current Psychometric Evaluation	3
2.0	METHOD	4
2.1	Sample Characteristics.....	4
2.2	Data Scrubbing.....	7
2.2.1	Missingness.....	7
2.3	Measures	8
2.4	Technical Approach.....	8
2.4.1	Technical Attachments.....	8
2.4.2	Statistical Methods.....	9
2.4.3	Terminology.....	14
2.5	Organization of the Report.....	14
3.0	RESULTS	14
3.1	Verbal Analogies	15
3.1.1	Descriptives.....	15
3.1.2	Difficulty.....	16
3.1.3	Discriminability	17
3.1.4	Internal Consistency if Item is Deleted.....	18
3.1.5	Omission Rates and Overt Guessing.....	19
3.1.6	Distractor Analysis.....	20
3.1.7	Common Items.....	20
3.1.8	Subgroup Differences	21
3.1.9	Conclusion	25
3.2	Arithmetic Reasoning	26
3.2.1	Descriptives.....	26
3.2.2	Difficulty.....	27
3.2.3	Discriminability	28
3.2.4	Internal Consistency if Item is Deleted.....	29
3.2.5	Omission Rates and Overt Guessing.....	30
3.2.6	Distractor Analysis.....	32
3.2.7	Common Items.....	32
3.2.8	Subgroup Differences	33
3.2.9	Conclusion	36
3.3	Word Knowledge	37
3.3.1	Descriptives.....	37
3.3.2	Difficulty.....	38
3.3.3	Discriminability	39
3.3.4	Internal Consistency if Item is Deleted.....	40
3.3.5	Omission Rates and Overt Guessing.....	41

3.3.6	Distractor Analysis.....	42
3.3.7	Common Items.....	42
3.3.8	Subgroup Differences	43
3.3.9	Conclusion	46
3.4	Math Knowledge.....	47
3.4.1	Descriptives.....	47
3.4.2	Difficulty.....	48
3.4.3	Discriminability	49
3.4.4	Internal Consistency if Item is Deleted.....	50
3.4.5	Omission Rates and Overt Guessing.....	51
3.4.6	Distractor Analysis.....	52
3.4.7	Common Items.....	52
3.4.8	Subgroup Differences	53
3.4.9	Conclusion	56
3.5	Reading Comprehension.....	57
3.5.1	Descriptives.....	57
3.5.2	Difficulty.....	58
3.5.3	Discriminability	59
3.5.4	Internal Consistency if Item is Deleted.....	60
3.5.5	Omission Rates and Overt Guessing.....	61
3.5.6	Distractor Analysis.....	62
3.5.7	Common Items.....	63
3.5.8	Subgroup Differences	63
3.5.9	Conclusion	66
3.6	Physical Science.....	67
3.6.1	Descriptives.....	67
3.6.2	Difficulty.....	68
3.6.3	Discriminability	69
3.6.4	Internal Consistency if Item is Deleted.....	70
3.6.5	Omission Rates and Overt Guessing.....	71
3.6.6	Distractor Analysis.....	72
3.6.7	Common Items.....	72
3.6.8	Subgroup Differences	73
3.6.9	Conclusion	76
3.7	Table Reading.....	77
3.7.1	Descriptives.....	77
3.7.2	Difficulty.....	79
3.7.3	Discriminability	80
3.7.4	Internal Consistency if Item is Deleted.....	82
3.7.5	Omission Rates and Overt Guessing.....	83
3.7.6	Distractor Analysis.....	84
3.7.7	Common Items.....	84
3.7.8	Subgroup Differences	85
3.7.9	Conclusion	91
3.8	Instrument Comprehension.....	91
3.8.1	Descriptives.....	91
3.8.2	Difficulty.....	92
3.8.3	Discriminability	94
3.8.4	Internal Consistency if Item is Deleted.....	94

3.8.5	Omission Rates and Overt Guessing.....	95
3.8.6	Distractor Analysis.....	96
3.8.7	Common Items.....	97
3.8.8	Subgroup Differences	98
3.8.9	Conclusion	101
3.9	Block Counting.....	102
3.9.1	Descriptives.....	102
3.9.2	Difficulty.....	103
3.9.3	Discriminability	105
3.9.4	Internal Consistency if Item is Deleted.....	106
3.9.5	Omission Rates and Overt Guessing.....	107
3.9.6	Distractor Analysis.....	108
3.9.7	Common Items.....	108
3.9.8	Subgroup Differences	109
3.9.9	Conclusion	113
3.10	Aviation Information	114
3.10.1	Descriptives.....	114
3.10.2	Difficulty.....	115
3.10.3	Discriminability	116
3.10.4	Internal Consistency if Item is Deleted.....	117
3.10.5	Omission Rates and Overt Guessing.....	118
3.10.6	Distractor Analysis.....	119
3.10.7	Common Items.....	119
3.10.8	Subgroup Differences	120
3.10.9	Conclusion	123
4.0	CONCLUSION.....	124
	REFERENCES	125
	APPENDIX A. SYMBOLS, ABBREVIATIONS AND ACRONYMS	128

LIST OF TABLES

Table 1. AFOQT Form T Subtests.....	1
Table 2. Unstratified Sample Demographics.....	5
Table 3. VA Unstratified Sample Descriptives	15
Table 4. VA Unstratified Sample Difficulty.....	16
Table 5. VA Unstratified Sample Discriminability	17
Table 6. VA Unstratified Sample Internal Consistency if Item is Deleted.....	18
Table 7. VA Unstratified Sample Omission Rates	19
Table 8. VA Unstratified Sample Common Items Difficulty	21
Table 9. VA Unstratified Sample Mean Score Subgroup Differences - Cohen's d	22
Table 10. VA Unstratified Sample DIF _{MH}	24
Table 11. VA Unstratified Sample DIF _{Lord}	25
Table 12. AR Unstratified Sample Descriptives.....	26
Table 13. AR Unstratified Sample Difficulty	27
Table 14. AR Unstratified Sample Discriminability	28
Table 15. AR Unstratified Sample Internal Consistency if Item is Deleted.....	29
Table 16. AR Unstratified Sample Omission Rates	31
Table 17. AR Unstratified Sample Common Items Difficulty	32
Table 18. AR Unstratified Sample Mean Score Subgroup Differences - Cohen's d	34
Table 19. AR Unstratified Sample DIF _{MH}	35
Table 20. AR Unstratified Sample DIF _{Lord}	36
Table 21. WK Unstratified Sample Descriptives.....	37
Table 22. WK Unstratified Sample Difficulty	38
Table 23. WK Unstratified Sample Discriminability	39
Table 24. WK Unstratified Sample Internal Consistency if Item is Deleted.....	40
Table 25. WK Unstratified Sample Omission Rates	41
Table 26. WK Unstratified Sample Common Items Difficulty	43
Table 27. WK Unstratified Sample Mean Score Subgroup Differences – Cohen's d	44
Table 28. WK Unstratified Sample DIF _{MH}	45
Table 29. WK Unstratified Sample DIF _{Lord}	46
Table 30. MK Unstratified Sample Descriptives.....	47
Table 31. MK Unstratified Sample Difficulty	48
Table 32. MK Unstratified Sample Discriminability.....	49
Table 33. MK Unstratified Sample Internal Consistency if Item is Deleted.....	50
Table 34. MK Unstratified Sample Omission Rates.....	51
Table 35. MK Unstratified Sample Common Items Difficulty	53
Table 36. MK Unstratified Sample Mean Score Subgroup Differences - Cohen's d	54
Table 37. MK Unstratified Sample DIF _{MH}	55
Table 38. MK Unstratified Sample DIF _{Lord}	56
Table 39. RC Unstratified Sample Descriptives.....	58
Table 40. RC Unstratified Sample Difficulty	59
Table 41. RC Unstratified Sample Discriminability.....	60
Table 42. RC Unstratified Sample Internal Consistency if Item is Deleted.....	61
Table 43. RC Unstratified Sample Omission Rates.....	62
Table 44. RC Unstratified Sample Mean Score Subgroup Differences - Cohen's d	64
Table 45. RC Unstratified Sample DIF _{MH}	65
Table 46. RC Unstratified Sample DIF _{Lord}	66
Table 47. PS Unstratified Sample Descriptives.....	67

Table 48. PS Unstratified Sample Difficulty	68
Table 49. PS Unstratified Sample Discriminability.....	69
Table 50. PS Unstratified Sample Internal Consistency if Item is Deleted.....	70
Table 51. PS Unstratified Sample Omission Rates.....	71
Table 52. PS Unstratified Sample Common Items Difficulty	73
Table 53. PS Unstratified Sample Mean Score Subgroup Differences - Cohen's d	74
Table 54. PS Unstratified Sample DIF_{MH}	75
Table 55. PS Unstratified Sample DIF_{Lord}	76
Table 56. TR Unstratified Sample Descriptives	78
Table 57. TR Unstratified Sample Difficulty	79
Table 58. TR Unstratified Sample Discriminability	81
Table 59. TR Unstratified Sample Internal Consistency if Item is Deleted	82
Table 60. TR Unstratified Sample Omission Rates.....	83
Table 61. TR Unstratified Sample Common Items Difficulty.....	85
Table 62. TR Unstratified Sample Mean Score Subgroup Differences - Cohen's d	87
Table 63. TR Unstratified Sample DIF_{MH}	89
Table 64. TR Unstratified Sample DIF_{Lord}	90
Table 65. IC Unstratified Sample Descriptives	92
Table 66. IC Unstratified Sample Difficulty	93
Table 67. IC Unstratified Sample Discriminability	94
Table 68. IC Unstratified Sample Internal Consistency if Item is Deleted	95
Table 69. IC Unstratified Sample Omission Rates	96
Table 70. IC Unstratified Sample Common Items Difficulty.....	97
Table 71. IC Unstratified Sample Mean Score Subgroup Differences - Cohen's d	99
Table 72. IC Unstratified Sample DIF_{MH}	100
Table 73. IC Unstratified Sample DIF_{Lord}	101
Table 74. BC Unstratified Sample Descriptives	103
Table 75. BC Unstratified Sample Difficulty	104
Table 76. BC Unstratified Sample Discriminability.....	105
Table 77. BC Unstratified Sample Internal Consistency if Item is Deleted	106
Table 78. BC Unstratified Sample Omission Rates.....	107
Table 79. BC Unstratified Sample Common Items Difficulty	109
Table 80. BC Unstratified Sample Mean Score Subgroup Differences - Cohen's d	110
Table 81. BC Unstratified Sample DIF_{MH}	112
Table 82. BC Unstratified Sample DIF_{Lord}	113
Table 83. AI Unstratified Sample Descriptives	114
Table 84. AI Unstratified Sample Difficulty	115
Table 85. AI Unstratified Sample Discriminability.....	116
Table 86. AI Unstratified Sample Internal Consistency if Item is Deleted	117
Table 87. AI Unstratified Sample Omission Rates.....	118
Table 88. AI Unstratified Sample Common Items Difficulty	119
Table 89. AI Unstratified Sample Mean Score Subgroup Differences - Cohen's d	121
Table 90. AI Unstratified Sample DIF_{MH}	122
Table 91. AI Unstratified Sample DIF_{Lord}	123

EXECUTIVE SUMMARY

The purpose of this technical report was to document the psychometrically evaluation of the items that made up the 10 cognitive subtests included in the Air Force Officer Qualifying Test (AFOQT) Form T. Below is a brief summary of the most notable findings between the two parallel forms (AFOQT Form T1 and T2) for the unstratified sample.

The items in the Verbal Analogies (VA) subtest had: (1) below average difficulty; (2) low discriminability; (3) slight speededness; (4) acceptable internal consistency; (5) two distractors chosen more frequently than the correct response; (6) some distractors chosen infrequently; and (6) a few large effect sizes for Black/African American and White subgroups.

The items in the Arithmetic Reasoning (AR) subtest had: (1) average difficulty; (2) low-to-moderate discriminability; (3) slight speededness; (4) acceptable internal consistency; (5) some distractors chosen more frequently than the correct response; and (6) one large effect size for Black/African American and White subgroup comparisons.

The items in the Word Knowledge (WK) subtest had: (1) below average difficulty; (2) moderate-to-high discriminability; (3) slightly high internal consistency suggesting redundancy; (4) one distractor chosen infrequently; and (5) some large effect sizes for Female and Male, Black/African American and White, and Asian and White subgroup comparisons.

The items in the Math Knowledge (MK) subtest had: (1) below average difficulty; (2) moderate-to-high discriminability; (3) slight speededness; (4) slightly high internal consistency suggesting redundancy; (5) one distractor chosen more frequently than the correct response; and (6) no large effect sizes across the subgroups.

The items in the Reading Comprehension (RC) subtest had: (1) dramatically below average difficulty; (2) low discriminability; (3) acceptable internal consistency; (4) many distractors chosen infrequently; and (5) a few large effect sizes for Black/African American and White subgroup comparisons. We recommend that the content be refreshed with new items/passages that would minimize vocabulary and cognitive load.

The items in the Physical Science (PS) subtest had: (1) average difficulty; (2) moderate discriminability; (3) acceptable internal consistency; (4) one distractor chosen infrequently; and (5) a few large effect sizes for female and male and Black/African American and White subgroup comparisons.

The items in the Table Reading (TR) subtest had: (1) below average difficulty; (2) moderate-to-high discriminability; (3) speededness; (4) slightly high internal consistency suggesting redundancy; (5) some distractors chosen more frequently than the correct response; (6) many distractors chosen infrequently; and (7) a few large effect sizes for Black/African American and White subgroup comparisons, likely due to speededness.

The items in the Instrument Comprehension (IC) subtest had: (1) below average difficulty; (2) high discriminability; (3) speededness (to a lesser extent than Table Reading or Block Counting); (4) high internal consistency suggesting redundancy; and (5) a few large effect sizes for

Black/African American and White subgroup comparisons. We recommend examining the item content for strategies to balance difficulty and mean score subgroup differences (e.g., increased time, practice items).

The items in the Block Counting (BC) subtest had: (1) below average difficulty; (2) high discriminability; (3) speededness; (4) slightly high internal consistency suggesting redundancy; (5) some distractors chosen more frequently than the correct response; (6) many distractors chosen more frequently than the correct response; (7) had many distractors chosen infrequently; and (8) no large mean score differences across the subgroups. We recommend reviewing the BC instructions to ensure that test-takers have practice items that reflect the items in the test.

The items in the Aviation Information (AI) subtest had: (1) average difficulty; (2) low-to-moderate discriminability; (3) acceptable internal consistency; (4) some distractors chosen more frequently than the correct response; and (5) no large mean score differences across the subgroups.

Based on this pattern of results, we recommend the following. First, most of the items seem to be of below average difficulty. The next iteration of the AFOQT should explicitly specify what level of difficulty would be preferred and develop new items or revise the existing items accordingly. Second, most of the items and their subtests seem to be of low or moderate discriminability. The items should be revised to differentiate between, say, low ability level and lower ability level (or, between high ability level and higher ability level). It is usually fairly easy for a test item to differentiate between high and low ability tests takers, however it is harder to design items that would differentiate between low and lower end of the ability spectrum. Third, many items between the parallel forms and across the subtests seem to have problematic distractors, such that the distractors are either chosen more frequently than the keyed response or are chosen infrequently. The problematic distractors should be revised. Fourth, based on the inflated Cronbach's alphas, some subtests may be shortened without sacrificing measurement precision (e.g., IC, TR). Finally, there were items with large mean score effect sizes (Cohen's *d*) and Differential Item Functioning (DIF) effect sizes across the subgroups. The items should be reviewed qualitatively to ensure that there are no obvious culturally-laden items. Any problematic items should be replaced to potentially reduce mean score subgroup differences.

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 items needs to be based on the holistic interpretation of the data resulting from: Subtest-level analyses (Walsh et al., 2022), composite-level analyses, criterion-related validities, adverse impact ^[1] 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 (SJT) (Walsh et al., 2021) and the Self-Description Inventory (Woolley et al., under review).

^[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 an item-level evaluation of the 10 cognitive subtests on the AFOQT Form T. This introductory section briefly overviews the history of the AFOQT, describes the composition of Form T, and describes 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 commissioning and aircrew training classification and is widely accepted among military personnel selection communities as a useful and cost-effective instrument. Historically, the AFOQT has been the primary selection test for the Air Force Reserve Officer Training Corps (AFROTC), Officer Training School (OTS), and the Airman Education and Commissioning Program (AECp). It is also used in the selection process for Undergraduate Pilot Training (UPT), Undergraduate Remotely Piloted Aircraft (RPA) Training (URT), Combat Systems 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 complete history of the AFOQT).

1.2 AFOQT Form T

The current version of the AFOQT (Form T) became operational on August 1, 2014. There are two parallel forms of the AFOQT Form T (identified as versions T1 and T2) to reduce practice effects and to preserve test security. The AFOQT is composed of 10 cognitive and two non-cognitive subtests and takes approximately 3.5 hours to administer. The subtests included in the AFOQT with their administration time and number of items are listed in Table 1.

Table 1. AFOQT Form T Subtests

Subtests (in the order of administration)	Subtest Abbreviation	Subtest Type	Subtest Length (in min)	Number of Items (Number of Response Options)	Number of Common Items
1. Verbal Analogies	VA	Cognitive	8	25 (5)	8
2. Arithmetic Reasoning	AR	Cognitive	29	25 (5)	8
3. Word Knowledge	WK	Cognitive	5	25 (5)	9
4. Math Knowledge	MK	Cognitive	22	25 (5)	9
5. Reading Comprehension	RC	Cognitive	38	25 (5)	0
6. Situational Judgment Test	SJT	Non-Cognitive	35	50 (5)	26
7. Self-Description Inventory - Officers	SDI-O	Non-Cognitive	45	240 (5)	240
8. Physical Science	PS	Cognitive	10	20 (5)	7
9. Table Reading	TR	Cognitive	7	40 (5)	10
10. Instrument Comprehension	IC	Cognitive	5	25 (4)	13
11. Block Counting	BC	Cognitive	4.5	30 (5)	15
12. Aviation Information	AI	Cognitive	8	20 (5)	7

Subtest scores are combined to generate composite scores used to predict success in certain types of the 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 AFOQT Form T 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

The participants included 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 States Air Force Academy (USAF), 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, Air Battle Manager (ABM), fighter or mobility pilot, and 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.

1.3 Previous 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 was evaluated by Carretta et al. (2016). A comparison between Form S and Form T was performed by Aguilar (2017). The current report compares the results obtained in the present psychometric investigation against the results observed in these three reports.

1.3.1 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 item-level psychometric properties in terms of item difficulty, overt guessing, internal consistency, and distractors. Some items were found to have moderate to large gender, racial, and ethnic mean score subgroup differences, low item-total correlations (ITCs), and poor distractors. The report contains item-level statistics.

1.3.2 Carretta et al. (2016)

Data for the Carretta et al. (2016) report consisted of responses from USAF officer applicants who were administered either AFOQT Form T1 ($N = 5,681$) or Form T2 ($N = 5,199$) between 2015 and

2016. In general, both forms demonstrated acceptable item-level psychometric properties within and between the parallel forms, with some items requiring further revisions. For example, there were some difficult items on BC which could be explained by the lack of clear instructions. Items tended to be omitted more frequently toward the end of each subtest indicating subtest speededness. Note that the authors performed item-level analyses for the Form T subtests, but the report contains only observations and trends observed at the subtest level. Therefore, the present study draws parallels with the study by Carretta et al. (2016) to the extent possible.

1.3.3 Aguilar (2017)

Using the same AFOQT Form T dataset as in Carretta et al. (2016) and the same AFOQT Form S dataset as in Morath et al. (2008), 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 (SDs) 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. Note that the author performed item-level analyses for the Form T subtests, but the report contains only observations and trends observed at the subtest level. Therefore, the present study draws parallels with the study by Aguilar (2017) to the extent possible.

1.3.4 Note on Form S and Form T

Note that there were some qualitative changes between Form S and Form T. Specifically, Form S included two spatial subtests (Rotated Blocks [RB] and Hidden Figures [HF]) that were not included in Form T. RC, which was included in Forms O through R, was not tested in Form S, but was added to Form T. The content of General Science (GS) in Form S was modified to focus on 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. In light of these considerations, the present report draws parallels between the findings observed in Form T and those reported by Morath et al. (2008) to the extent possible.

In comparing the results of the present study for Form T against the studies by Aguilar (2017) and by Morath et al. (2008) for Form S, it is important to note that the common items within the forms are not always the same as between the forms. This is described in more detail in Section 2.4.2 under the ‘Common Items’ subheading.

1.3.5 The Current Psychometric Evaluation

The present study was initiated by the Air Force Personnel Center Strategic Research and Assessments Branch (AFPC/DSYX), who after collecting a substantial amount of operational data was interested in evaluating the psychometric integrity of the cognitive and non-cognitive subtests of the AFOQT Form T at the item-, subtest-, and composite-levels and to make recommendations for potential changes and enhancements to future revisions of the test. Other technical reports describe the details of the subtest-level and composite-level analyses performed on the cognitive

subtests; and of the item-level, subtest-level, and composite-level analyses performed on the non-cognitive subtests (i.e., Kantrowitz et al., in press; Walsh, Woolley, Brady, & Carretta, 2022; Walsh, Woolley, Brady, Melick, & Carretta, 2021; Woolley et al., in progress). This technical report serves as the final deliverable for the item-level analyses performed on the 10 cognitive subtests.

2.0 METHOD

This section describes the methodology of the item analyses, describes the sample used in these analyses, and outlines the organization of the report.

2.1 Sample Characteristics

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

Table 2 shows sample sizes (*N*) and proportions (%) of each demographic category in Forms T1 and T2 for first-time test-takers. Note that when a demographic is reported as ‘Unknown,’ it indicates that an examinee passively declined to respond or that their response was not legibly marked (i.e., did not select a response). ‘Declined to Respond’ indicates that a test-taker actively declined to provide a response (i.e., selected “Decline to Respond”).

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	1751	5%	1,706	5%
Somewhat higher than average	10197	28%	9,421	29%
Average	14079	39%	12,543	38%
Somewhat lower than average	6216	17%	5,594	17%
Much lower than average	2113	6%	2,009	6%
Declined to respond	1629	5%	1,576	5%
Unknown	114	0%	103	0%
Years of Education Completed				
12	1177	3%	1,251	4%
13	10382	29%	9,994	30%
14	4847	13%	4,314	13%
15	3863	11%	3,368	10%
16	10790	30%	9,441	29%
17	2294	6%	2,068	6%
18	1849	5%	1,697	5%
19	399	1%	359	1%
20	218	1%	204	1%
21+	221	1%	190	1%
Unknown	59	0%	66	0%
Academic Degree				
High School Diploma	17420	48%	16,359	50%
Associate's Degree	3741	10%	3,393	10%
Bachelor's Degree	13023	36%	11,402	35%
Master's Degree	1692	5%	1,628	5%
Doctoral Degree	128	0%	89	0%
Unknown	95	0%	81	0%
Accession Source				
OTS-CIV	8026	22%	7,284	22%
OTS-AD	5430	15%	4,843	15%
AACP	170	0%	168	1%
USAF	4312	12%	3,741	11%
ROTC	12765	35%	12,413	38%
ANG	3915	11%	3,288	10%
AFRES	1481	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. N_{T1} = 36,099 and N_{T2} = 32,952.

As can be seen, the demographic composition of the two samples was similar. The mean age of T1 test-takers was 23.79 ($SD = 5.04$). The mean age of T2 test-takers was 23.65 ($SD = 5.00$). The mean education levels were 14.88 years for T1 and 14.82 years for T2. Both samples predominantly consisted of Male (74% for Forms T1 and T2) and White (77% for Forms T1 and T2) test-takers.

Regarding gender, demographics varied across the accession sources. There were fewer female examinees in OTS-AD (T1: 21%; T2: 19%) compared to OTS-CIV (T1: 30%; T2: 29%) and ROTC (T1: 28%; T2: 29%). However, these values are on par with the general USAF applicant test-taking gender distribution, with females making up roughly 30%.

Regarding racial demographic categories, there were fewer Black/African American examinees in USAFA (T1: 10%; T2: 9%) and more in AFRES (T1: 20%; T2: 19%) compared to the unstratified sample. There are fewer Asian test-takers in ANG (T1: 7%; T2: 7%), ROTC (T1: 12%; T2: 11%), and USAFA (T1: 11%; T2: 11%) compared to the unstratified sample.

Consistent with previous efforts, the present analyses did not focus on American Indian/Alaska Native, Native Hawaiian/Other Pacific Islander, or Multiracial subgroups due to their relatively small sample sizes.

Regarding ethnic categories, there were fewer Hispanic examinees in USAFA (T1: 11%; T2: 11%) and in ANG (T1: 11%; T2: 10%) than in the unstratified sample. There are more Hispanic examinees in ROTC (T1: 16%; T2: 17%) and in AFRES (T1: 16%; T2: 15%) than in the unstratified sample. Note that the AECP population is fairly small, with some demographics having less than 10 people.

Consistent with these demographic differences across stratifications, Morath et al. (2008) found that in Form S Black/African American, 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.

Note that throughout the technical report, female examinees will be referred to as either female or F; male examinees will be referred to as either male or M; Black/African American examinees will be referred to as either Black/African American, Black, or B; White examinees will be referred to as either White or W; Asian examinees will be referred to as either Asian or A; American Indian/Alaska Native examinees will be referred to as either American Indian/Alaska Native or AIAN; Native Hawaiian/Other Pacific Islander examinees will be referred to as either Native Hawaiian/Other Pacific Islander or NHPI; White Hispanic examinees will be referred to as either White Hispanic or WH; White Non-Hispanic examinees will be referred to as either White Non-Hispanic or WnH; Black/African American Hispanic examinees will be referred to as either Black Hispanic or BH; finally, Black/African American Non-Hispanic examinees will be referred to as either Black Non-Hispanic or BnH.

2.2 Data Scrubbing

AFPC/DSYX performed the initial data scrubbing. Specifically, cases were deleted listwise if: (1) they had an invalid social security number (e.g., 123-45-6789); (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 examinee was assigned a unique alpha-numeric identifier.

The initially scrubbed dataset from AFPC/DSYX included 78,397 cases across Forms T1 and T2. Note that several items were discovered to have been mis-keyed – T1 PS Item 13, 16, and 17; T2 PS Item 14; and T2 RC Item 12. These items were re-keyed. Examinees who self-identified as belonging to the Test Control Officer-Test Evaluator, AD-Other, or CIV-Other accession sources were removed listwise 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-level scores. Then, unit missingness was removed listwise. Section 2.2.1 describes how we handled missingness in more detail ^[2].

2.2.1 Missingness

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 demographics, but no responses to the items 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). For a complete explanation of each of these terms, please refer to Schafer and Graham (2002). In brief, *MCAR* means that a missing response on variable X for an examinee does not

^[2] Note that in the subsequent effort involving the subtest-level analyses, careless responding was detected at the item level (also known as Insufficient Effort Responding [IER]). Unit IER were removed listwise at that time. Unit IER was not removed from this research and ranges from 0 to 24 cases on T1 and from 2 to 26 cases on T2.

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 examinees). *MAR* means that the missing response on Variable X may 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 revealed that the percentage of missing data increased for items positioned later in their respective subtest (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). Items from 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, these subtests were designed to be speeded and therefore missingness simply indicates that the examinees ran out of time.

2.3 Measures

The constructs that the 10 cognitive subtests were designed to measure are listed below (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 attitude 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.

2.4 Technical Approach

The analyses were restricted to first-time test-takers. Note that throughout this report, the discussion involves the unstratified sample, which includes test-takers across all accession sources, and the stratified samples, which include test-takers from each of the seven accession sources: (1) OTS-CIV, (2) OTS-AD, (3) AECP, (4) USAFA, (5) ROTC, (6) ANG, and (7) AFRES.

2.4.1 Technical Attachments

Note that this technical report refers to three attachments. Attachment 1 (titled “Cognitive Items Summary”) is an Excel workbook containing the following: (1) Forms T1 and T2 demographics information for stratified and unstratified samples; (2) Forms T1 and T2 Classical Test Theory (CTT) results for stratified and unstratified samples; (3) Forms T1 and T2 Item Response Theory (IRT) results for stratified and unstratified samples; (4) common items mapping and results for

stratified and unstratified samples; and (5) a summary of problematic items. Attachment 2 (titled “Q&A”) is a word document containing more in-depth coverage of the questions raised during the execution of this project and the IST team’s detailed answers to those questions (e.g., What Type I error correction methods were used for DIF analyses? How are Form S and Form T different in terms of difficulty across the accession sources?). Attachment 2 also contains the Principal Components Analysis (PCA) scree plots that were generated to establish sufficient unidimensionality of the subtests to permit the IRT analyses. Attachment 3 (title “Cohen’s d ”) contains a short overview of the conventional cutoffs for Cohen’s d effect sizes and the cutoffs adopted in the current research.

2.4.2 Statistical Methods

Both CTT and IRT approaches were used in item evaluation. Items were evaluated holistically, based on the pattern of observations across the different types of analyses, and taking into account the strengths and weaknesses of each analysis. The following paragraphs describe the statistics that were generated.

Scoring

Item scoring was conducted after data scrubbing was completed (section 2.2 describes the data scrubbing steps in more detail). Items were scored by applying a scoring key provided by AFPC/DSYX. Items were scored dichotomously (0 = incorrect or missing; 1 = correct).

CTT

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

Descriptives. The first step involved the examination of the item descriptive statistics, including means (M), SDs, skewness, and kurtosis. These descriptives, also known as measures of central tendency, help identify poor performing or well performing items. “Generally, the higher the variability of the items and the more the mean of the item is at the center point of the distribution, the better the item will perform” (Kline, 2005, p. 95). Recall that the items were scored dichotomously and therefore the items’ means were also the items’ difficulty parameters (p -values, described in more detail below). Negative skewness indicates that the majority of the scores are

clustered around the upper end of the distribution and therefore mark easier items (Tabachnick et al., 2007). Positive skewness indicates that the majority of the scores are clustered around the lower end of the distribution and therefore mark more difficult items (Tabachnick et al., 2007). The skewness of 0 indicates a normal distribution. Skewness at or below -3 and at or above 3 is considered extreme (Kline, 2011). Positive kurtosis indicates heavy tails and peakedness relative to the normal distribution, whereas negative kurtosis indicates light tails and flatness (DeCarlo, 1997). Most studies do not interpret and some do not report kurtosis. With the larger sample sizes such as in this study, the impact of negative or positive kurtosis is diminished (Tabachnick et al., 2007).

Difficulty and Speededness. Item difficulty and speededness were assessed by observing p -values, omission rates, and overt guessing rates. The p -values represent the percentage of examinees who endorsed the correct response option, and as mentioned above, also represent an item's mean (Kline, 2005). High p -values indicate easy items and low p -values indicate difficult items (Kline, 2005). In general, items that have difficulty of 0 or 1 are useless, because they do not provide variability (information) about examinees. More specifically, items with p -values at or below .20 for all subtests except IC and at or below .25 for IC are considered too difficult and represent a floor effect. Items with p -values at or above .80 for all subtests except IC and at or above .75 for IC are considered too easy and represent a ceiling effect. Literature suggests that items with p -values of .50 are optimal, because they provide the best differentiation between low-ability and high-ability test-takers (Kline, 2005). However, the evaluation of whether an item is difficult or easy should be based on the goals of the selection instrument.

To examine omission rates, we looked at (1) each item (percentage of examinees who skipped an item); (2) a mutually exclusive grouping of five items (percentage of examinees who skipped at least one item out of a group of five items); and (3) longstring for the last five items (percentage of test-takers who skipped all last five items). When evaluating omission rates, it is important to look at the subtests holistically to observe trends. If the omissions increase throughout the subtest, it may indicate that test-takers skip items because they run out of time (evidence of speededness), because the items are difficult, or a combination of the two.

To examine overt guessing rates, we looked at the percentage of examinees who systematically endorsed the same response option for the last five items (i.e., A, A, A, A, A). This pattern of responding may indicate that test-takers bubble in their Scantrons with the same response options hoping to guess some items right when the time runs low (evidence of speededness).

Item Discriminability. Item discriminability, expressed as an ITC, refers to a correlation between each item and the total subtest score (computed with the item in question removed; Kline, 2005). Much like the Pearson product moment correlation, ITC's range from -1 to 1, where values closer to 1 are considered strong and values closer to 0 are considered weak. Negative values indicate that the item is negatively related to the subtest score, which is highly undesirable. None of the items across the cognitive subtests had negative ITC's.

Internal Consistency if Item is Deleted. Internal consistency at the subtest level, expressed as Cronbach's alpha, concerns the interrelatedness of items within a subtest (Schmitt, 1996). While the subtest-level alphas are discussed in greater detail elsewhere (see Walsh et al., 2022), the

current report provides subtest-level alphas for reference. Internal consistency at the item level shows how the alpha would change if an item was deleted (Schmitt, 1996). Items that contribute to an upward increase in Cronbach's alpha may need to be revised or deleted from the subtest. Generally, alphas below .70 are considered low, between .70-.80 acceptable, and above .80 high (some might argue these alphas might indicate redundancy in measurement such that a reduction of items may improve the measure; Cronbach, 1951; Schmitt, 1996).

Distractor Analysis. Distractors refer to incorrect response options that may appear as correct (Thissen et al., 1989). An effective distractor should attract at least some test-takers. Problematic distractors might attract (1) high-ability test-takers or (2) no test-takers. For each item, we calculated the frequencies of the endorsement rate across all response options. Of particular concern were distractors that were selected more frequently than the correct answer and the response options that were selected 1% of the time or less.

Item Mean Score Subgroup Differences. For each item, we examined item mean score differences between the majority subgroups and legally protected minority subgroups. These differences are estimated using Cohen's d , which is a measure of mean score differences among the two groups expressed in standard deviation units, also known as an effect size (Cohen, 1992).

Conventionally, the magnitudes of .20 would be considered small; .50 would be considered moderate; and .80 or greater would be considered large (Cohen, 1992). The USAF traditionally has interpreted the magnitudes slightly differently (Shore et al., 2020). Specifically, "[T]he Air Force defines the threshold of concern for adverse impact^[3] as when subgroup mean score difference (Cohen's d) reach .40 (Shore et al., 2020, p. 10). "Any discussion of the existence of adverse impact for a measure refers to whether the effect size crosses the threshold of $d = 0.40$ " (Shore et al., 2020, p. 44). As a compromise between the usual conventions and the most recent studies in the USAF, this report treated Cohen's d of $\leq .40$ as small (likely not leading to adverse impact); Cohen's d between .40 and .80 as moderate (likely leading to adverse impact); and Cohen's $d \geq .80$ as large (highly likely leading to adverse impact). See Attachment 3 for more information. Although these cutoffs were slightly different than the ones prescribed by the literature, they are in line with the cutoffs used for the Armed Services Vocational Aptitude Battery (ASVAB) and other standardized assessments throughout the Department of Defense (DoD). Moderate-to-large effect sizes suggest that an item is biased against one group over the other and may contribute to adverse impact. It is important to note that when discussing Cohen's d for the stratified samples, the sample sizes for AECF were small across the aforementioned subgroups and therefore the results are to be interpreted with caution.

Common Items. Common items are items that are identical between Forms T1 and T2. As Table 1 shows, the number of common items varied among the subtests, ranging from 0 for RC to 15 for BC. The test developers made an effort to place the common items in similar locations in the

^[3] 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).

parallel forms (e.g., VA T1-10 is common with VA Item T2-11). These items should have similar psychometric properties between Forms T1 and T2. If there are differences, it may indicate that the two forms are not parallel, which may require equating, or it may indicate significant differences in samples between Forms T1 and T2. To evaluate the performance of the common items between the two forms, *p*-values were compared with independent samples *t*-tests. Specifically, *t*-test significant levels and effect sizes (Cohen's *d*) were examined, using the same cutoffs as in the item mean score subgroup difference analyses.

In comparing the results of the present study for Form T against the studies by Aguilar (2017) and by Morath et al. (2008) for Form S, it is important to note that the common items within the forms are not always the same as between the forms. Traditionally, the AFOQT form developers kept some items common between the forms (e.g., Form O and Form P) to allow for equating, but included additional items that were common within the parallel versions of the same form (e.g., Form P1 and P2). For example, there were 12 common items in the VA subtest between S1 and S2. There were eight common items in the same subtest between Forms T1 and T2. Only two items on S1 were common with T1 and three items on S2 were common with T2. The placement of the items between the forms is not stable either. For example, T1-13 is common with S2-07. Given that many subtests are speeded, it is difficult to compare common items between the forms. Therefore, when making a comparison in this report against other reports, we simply look at the trend, not at any specific common items between the forms.

IRT

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

In applying IRT to the cognitive subtests, it was decided to fit three dichotomous IRT models (1PL, 2PL, and 3PL) to the data, because (1) the items were scored as '0' and '1' and (2) the results of the PCA revealed sufficient unidimensionality for each subtest. While a detailed description of each of the dichotomous models is beyond the scope of the current report, suffice it is to say that each model generates a different number of parameters. The 1PL model generates one parameter (difficulty); the 2PL model generates two parameters (difficulty and discriminability); and the 3PL model generates three parameters (difficulty, discriminability, and pseudo-guessing). After fitting these models, it was concluded that the 2PL model fit the data better, because it provides slightly more information than the 1PL model and allows for convergence, which was not achieved for some subtests when fitting the 3PL model (especially for TR and BC for unstratified and stratified samples; most of the subtests for the AECP sample). Therefore, the discussion in this report will be limited to the 2PL model output. The output for the other two models is provided in Attachment 1 but is not discussed in the report. For general information on the IRT models and frameworks, please refer to Lord and Novick (2008), Reise and Henson (2003), or Zickar (2012).

2PL b Parameter. As mentioned above, the 2PL model generated two parameters – difficulty and discriminability. Difficulty parameter (*b*) describes a point on the latent trait continuum where the probability for endorsing a correct response .5 (Reise & Waller, 2002). For example, a 2PL *b* parameter equal to -1.29 for VA Item T1-01 would indicate that test-takers with a latent trait standing lower than -1.29 would likely get the item wrong; test-takers with a latent trait standing higher than -1.29 would get the item right; and test-takers with a latent trait of exactly -1.29 would have a 50% chance of getting the item right. Depending on the goals of a psychological instrument, it might be desired to target certain ranges of the latent trait. Therefore, the items with the 2PL *b* parameter within the desired range should provide the most amount of information. For example, if the desire to filter out test-takers with lower standing on verbal ability, items that make up a subtest designed to measure the construct should target latent trait range within two *SD* below the mean (the items would be fairly easy). Applying the literature recommendation to the AFOQT (select-out strategy), 2PL *b* parameters below two standard deviations of the mean would be considered too easy; between two standard deviations of the mean and the mean would be considered acceptable; and above the mean would be considered difficult.

2PL a Parameter. Discriminability parameter (*a*) describes the extent to which the item is capable of differentiating between test-takers with lower and higher standing on the latent trait (Masters & Wright, 1997). Generally, 2PL *a* parameters below 1.00 are considered as indicators of poorly discriminating items and 2PL *a* parameters at or above 1.00 are considered as indicators of well discriminating items (Reise & Waller, 2002). 2PL *a* parameters below 0 are usually a sign of a poor convergence between the IRT model and the data. As mentioned before,

DIF. Subgroup differences were also examined using DIF. DIF exists when individuals with equal trait levels but from different subgroups perform differently on the same test item (Stark et al., 2004).

Literature shows that DIF detection is influenced by factors like sample size, test length, and group differences. Regarding *sample size*, research shows that large sample size (up to 1,500) and equal ratio of sample size tends to inflate the Type I error rates for parametric and non-parametric DIF methods (Kim, 2010). Regarding *test length*, both parametric and non-parametric DIF methods performed similarly in the longer tests (i.e., 40 items). In the shorter tests (i.e., 20 items), however, the Type I error rates of the non-parametric methods were higher than those of the parametric methods. Also, the effect of the test length on the Type I error rates of the parametric methods were significant: a longer test (e.g., 40 items test) inflated the Type I error rate more than a shorter test (e.g., 20 items test). The effect of the test length was not significant for the non-parametric methods. Regarding *group differences*, when the condition of the group difference was concerned, the presence of the group mean difference did not influence the Type I error rates much. However, the presence of the group *SD* difference significantly inflated the Type I error rates of parametric and non-parametric DIF methods.

There are many DIF methods and Type I error adjustment methods in the case of multiple significance testing (Kim, 2010). There are two main types of DIF methods: IRT and non-IRT. Non-IRT methods use test total scores as estimates of participant's trait value, as is done in CTT. IRT methods use a latent variable framework to estimate trait values directly. Within these two

types of DIF methods, there are many specific methods that can be used. Each method has been shown to have advantages over the other methods under certain conditions, such as sample size, test length, and subgroup differences (Gao, 2019). There are two main types of DIF: Uniform and nonuniform. Uniform DIF exists when item difficulty is found to differ across groups. Nonuniform DIF exists when item discrimination differs across groups. In this study, we used Mantel-Haenszel (MH) uniform approach and Lord uniform approach, both with Benjamini-Hochberg correction for alpha inflation. In general, absolute values of effect sizes of less than 1 are considered small, between 1 and 1.5 – moderate, and higher than 1.5 – large. With DIF being susceptible to Type I error due to large sample sizes, it is better to look for consistencies between and within forms when interpreting the DIF effect sizes to avoid capitalizing on chance. For a more detailed discussion, please refer to Attachment 2.

Item Drift. Item drift refers to changes in item parameters over time due to a variety of reasons, including breach of test security (Chan et al., 1999). Item drift can be assessed with DIF, using the same guidelines as for the subgroup differences. To examine item drift, the dataset was split into an early group (test-takers who took the AFOQT between January 2015 and August 2016) and a later group (test-takers who took the AFOQT between May 2018 and Dec 2019). The examination of the effect sizes comparing the early group and the later group revealed no evidence for item-level test security breach across the items. Therefore, the results for the item drift analyses are not discussed in any further detail in the remainder of the report.

2.4.3 Terminology

Throughout the report, references will be made to specific items in the subtests. To identify the items, we adopted a nomenclature where the first two characters denote either Form T1 or T2 and the other two digits denote a location of the item within the subtest. The order of the items resets for each subtest to start with 1. For example, an item referenced as T1-02 denotes a second item on T1.

2.5 Organization of the Report

This report is organized by chapters, with each chapter devoted to one of the 10 cognitive subtests. Each chapter describes the pattern of CTT and IRT findings for the stratified and unstratified samples and compares the results against the three aforementioned technical reports. At the end of each chapter, we provide a conclusion and recommendations for the subtest.

3.0 RESULTS

The following sections describe the results of the item analyses for each of the 10 cognitive subtests.

3.1 Verbal Analogies (VA)

VA is the first cognitive subtest presented to test-takers and is one of the five subtests that make up the Officership half of the AFOQT. It consists of 25 items with five multiple-choice options. Test-takers are allotted 8 minutes to complete the test, for an average of 19.2 seconds per item.

3.1.1 Descriptives

Table 3 summarizes the means, standard deviations, skewness, and kurtosis for the unstratified sample across Forms T1 and T2. The items means ranged from .31 to .80 with an average of .59 on T1 and from .32 to .83 with an average of .58 on T2. The means generally decreased throughout the test. The means for the first five items exceeded .60, whereas the means for the last five items fell below .50 across Forms T1 and T2. Skewness tended to switch from negative to positive throughout the test, with the values remaining within the acceptable range.

Table 3. VA Unstratified Sample Descriptives

Item #	T1				T2			
	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
01	.77	.42	-1.27	-.38	.72	.45	-1.00	-1.00
02	.74	.44	-1.12	-.74	.83	.37	-1.79	1.22
03	.62	.49	-.49	-1.76	.77	.42	-1.28	-.36
04	.80	.40	-1.52	.30	.66	.47	-.68	-1.54
05	.74	.44	-1.10	-.80	.68	.47	-.77	-1.41
06	.63	.48	-.54	-1.71	.66	.47	-.68	-1.54
07	.65	.48	-.64	-1.59	.64	.48	-.58	-1.67
08	.70	.46	-.87	-1.25	.61	.49	-.44	-1.81
09	.61	.49	-.45	-1.80	.59	.49	-.37	-1.86
10	.62	.49	-.49	-1.76	.63	.48	-.54	-1.71
11	.64	.48	-.60	-1.64	.59	.49	-.37	-1.86
12	.62	.48	-.50	-1.75	.71	.45	-.91	-1.16
13	.73	.44	-1.06	-.87	.57	.49	-.29	-1.92
14	.69	.46	-.82	-1.33	.63	.48	-.53	-1.72
15	.62	.49	-.50	-1.75	.72	.45	-.96	-1.08
16	.61	.49	-.45	-1.80	.48	.50	.09	-1.99
17	.50	.50	.01	-2.00	.55	.50	-.22	-1.95
18	.63	.48	-.53	-1.71	.60	.49	-.40	-1.84
19	.55	.50	-.20	-1.96	.50	.50	.01	-2.00
20	.50	.50	.02	-2.00	.40	.49	.41	-1.83
21	.35	.48	.62	-1.62	.48	.50	.07	-1.99
22	.40	.49	.39	-1.85	.48	.50	.09	-1.99
23	.39	.49	.46	-1.78	.32	.47	.76	-1.42
24	.31	.46	.83	-1.32	.41	.49	.38	-1.86
25	.31	.46	.84	-1.30	.36	.48	.60	-1.64

Note. *M* = Mean; *SD* = Standard Deviation. $N_{T1} = 36,098$; $N_{T2} = 32,950$.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing means and skewness switching from negative to positive throughout the test. USAFA cadets had the highest average mean compared with the other accession sources, with T1 averaging

at .63 and T2 averaging at .64. Carretta et al. (2016), Aguilar (2017), and Morath et al. (2008) did not report item-level central tendency statistics.

3.1.2 Difficulty

Table 4 summarizes the *p*-values and 2PL *b* parameters for the unstratified sample across Forms T1 and T2. The *p*-values ranged widely on both Forms T1 and T2, but in general, they tended to decrease throughout the test. Potential ceiling effects were observed for one item in T1 (T1-04: *p*-value = .80) and for one item in T2 (T2-02: *p*-value = .83). No items reached a floor effect. The 2PL *b* parameters tended to switch from negative to positive throughout the test, with most items staying within the acceptable range. The three exceptions were T1-04, T1-25, and T2-03, for which the 2PL *b* parameters were below the +/-2 threshold.

Table 4. VA Unstratified Sample Difficulty

Item #	T1		T2	
	<i>p</i> -values	2PL <i>b</i>	<i>p</i> -values	2PL <i>b</i>
01	.77	-1.29	.72	-1.19
02	.74	-1.29	.83	-1.56
03	.62	-.52	.77	-2.11
04	.80	-2.77	.66	-.66
05	.74	-1.81	.68	-1.06
06	.63	-.67	.66	-.59
07	.65	-.70	.64	-.72
08	.70	-1.47	.61	-.46
09	.61	-.71	.59	-.39
10	.62	-.56	.63	-.72
11	.64	-.79	.59	-.41
12	.62	-.82	.71	-1.23
13	.73	-1.37	.57	-.55
14	.69	-1.41	.63	-.92
15	.62	-.77	.72	-1.64
16	.61	-1.42	.48	.17
17	.50	.01	.55	-.54
18	.63	-.73	.60	-.86
19	.55	-.36	.50	.01
20	.50	.02	.40	.76
21	.35	.92	.48	.11
22	.40	.46	.48	.13
23	.39	.63	.32	1.44
24	.31	1.64	.41	.65
25	.31	2.13	.36	1.06

Note. *p*-value = Difficulty; 2PL *b* = IRT Difficulty Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing *p*-values and 2PL *b* parameters switching from negative to positive throughout the test. From the CTT perspective, USAFA cadets scored higher on average than the rest of the accession sources, with T1 averaging at .63 and T2 averaging at .64. No items reached a floor effect. There were some items across all stratifications that reached ceiling effect, with USAFA having the most

items (three on T1 and four on T2). Notably, all items with ceiling effect were towards the beginning of the test across the accession sources. From the IRT perspective, there were some items across all the accession sources that had at least one 2PL b parameter below the acceptable threshold. OTS-CIV and ROTC on T2 were the only exception to this, with all items showing acceptable 2PL b parameters.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and by Aguilar (2017) for Form T. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seemed to be more difficult. The mean p -values for Forms T1 and T2 were .59 and .58, respectively. The mean p -values for Form S1 and S2 were .66 and .65, respectively.

3.1.3 Discriminability

Table 5 summarizes the ITC's and 2PL a parameters for the unstratified sample across Forms T1 and T2. ITC's ranged from .24 to .48 with an average of .37 for T1, and from .29 to .54 with an average of .38 on T2. Five items on T1 and six items on T2 had acceptable 2PL a parameter values.

Table 5. VA Unstratified Sample Discriminability

Item #	T1		T2	
	ITC	2PL a	ITC	2PL a
01	.43	1.17	.41	.97
02	.40	.99	.41	1.35
03	.48	1.19	.30	.62
04	.26	.54	.51	1.38
05	.32	.63	.38	.80
06	.42	.95	.54	1.70
07	.46	1.12	.42	.94
08	.32	.62	.49	1.26
09	.37	.70	.48	1.24
10	.46	1.09	.41	.85
11	.41	.87	.47	1.17
12	.35	.66	.37	.82
13	.38	.85	.33	.56
14	.33	.61	.35	.62
15	.38	.71	.32	.61
16	.24	.32	.33	.55
17	.35	.61	.29	.41
18	.40	.84	.31	.49
19	.35	.60	.33	.52
20	.38	.72	.33	.58
21	.37	.74	.39	.71
22	.45	1.00	.40	.75
23	.40	.84	.32	.55
24	.30	.52	.34	.63
25	.26	.40	.34	.61

Note. ITC = Item-Total Correlation; 2PL a = IRT Discriminability Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample, such that there was a wide range of ITC's among the items and the majority of the items had low 2PL a parameters. Notably, OTS-AD and USAFA had no items with acceptable 2PL a parameters for T1, while other accession sources had at least four items with acceptable 2PL a parameters. T2 showed a better pattern, with all accession sources having at least five items with acceptable 2PL a parameters.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and by Aguilar (2017) for Form T. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seemed to be more discriminable ($T1 = .37$ and $T2 = .38$ vs. $S1 = .27$ and $S2 = .24$).

3.1.4 Internal Consistency if Item is Deleted

Table 6 summarizes Cronbach's alphas if an item is deleted for the unstratified sample across Forms T1 and T2. For reference, the VA subtest-level alphas for T1 and T2 were .74 and .76, respectively. The item-level alphas were high, ranging from .72 to .74 on T1 and from .74 to .76 on T2. The removal of any one of the items would have a small effect on the internal consistency of the subtest.

Table 6. VA Unstratified Sample Internal Consistency if Item is Deleted

Item #	T1 α	T2 α
01	.73	.75
02	.73	.75
03	.72	.75
04	.74	.74
05	.73	.75
06	.73	.74
07	.72	.75
08	.73	.74
09	.73	.74
10	.72	.75
11	.73	.74
12	.73	.75
13	.73	.75
14	.73	.75
15	.73	.75
16	.74	.75
17	.73	.76
18	.73	.75
19	.73	.75
20	.73	.75
21	.73	.75
22	.72	.75
23	.73	.75
24	.74	.75
25	.74	.75

Note. α = Cronbach's Alpha.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the alphas were all within the acceptable range and no one item was indicated as dramatically improving internal consistency if deleted. The USAFA sample was the only exception to this, with all items being below the .70 threshold on Forms T1 and T2.

Although the reports by Aguilar (2017), Carretta et al. (2016), and Morath et al. (2008) did not report the values for alpha if items were deleted, the Cronbach's alpha they reported for the subtests were similar to the alphas observed in this research.

3.1.5 Omission Rates and Overt Guessing

Table 7 summarizes the percentage of omissions in a form of missing values for each item and for a grouping of five items for Forms T1 and T2. The longstring omission rate across the last five items was 2.24% for T1 and 2.46% for T2. In general, the omission rates increased throughout the test. On T1, the first five items had an average omission rate of .37% whereas the last five items had an average rate of 6.13%. On T2 these respective rates were .32% and 5.45%. The longstring overt guessing for the last five items was 1.89% for T1 and 1.56% for T2.

Table 7. VA Unstratified Sample Omission Rates

Item #	T1		T2	
	For Each Item	Across Five Items	For Each Item	Across Five Items
01	.35%	.37%	.29%	.32%
02	.35%		.12%	
03	.32%		.38%	
04	.15%		.45%	
05	.69%		.33%	
06	.38%	.45%	.48%	.62%
07	.40%		.34%	
08	.19%		.69%	
09	.40%		1.24%	
10	.86%		.35%	
11	.55%	.64%	1.25%	.79%
12	.48%		.28%	
13	.31%		.38%	
14	1.04%		.79%	
15	.84%		1.23%	
16	2.61%	2.12%	.51%	1.80%
17	1.26%		3.23%	
18	1.89%		1.49%	
19	1.52%		1.76%	
20	3.30%		2.02%	
21	3.06%	6.13%	3.61%	5.45%
22	4.36%		3.93%	
23	7.00%		4.95%	
24	7.25%		7.20%	
25	9.01%		7.58%	

Note. For Each Item = Percent of missing values in each item; Across Five Items = Percent of missing values across a grouping of five items.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the omission rates tended to increase throughout the test. The accession sources varied in their omission rates across the last five items on both Forms T1 and T2. The lowest and the highest omission rates on T1 were observed for AECP (3.88%) and USAFA (8.30%), respectively. On T2, the lowest omission rate was observed for ROTC (4.73%) and the highest omission rate was observed for AECP (8.10%).

In general, the current pattern of findings is consistent with the results reported by Morath et al. (2008) for Form S and by Carretta et al. (2016) for Form T for the unstratified sample. These reports did not include the omission and guessing rates for the stratified samples. Aguilar (2017) did not report omission rates or overt guessing.

3.1.6 Distractor Analysis

When examining the response frequencies, the correct response was chosen more frequently than a distractor for all items on both Forms T1 and T2, except for T1-24 and T2-23. The items for which at least one of the distractors was chosen less than 1% of the time were: T1-03, T1-06, T2-03, T2-12, T2-16, and T2-18.

The pattern of results for the stratified samples was similar to the unstratified sample, such that all but two accession sources had at least one item for which a distractor was chosen more often than the correct response. The exceptions were ANG and AFRES on T2, with no distractors being chosen more often than the correct response. There were some items for which at least one distractor was chosen less than 1% of the time (e.g., T1-03 for OTS-CIV; T2-01 for OTS-AD).

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) for Form T. Aguilar (2017) did not report the distractor analysis. When comparing the present results against the results for Form S reported by Morath et al. (2008) for the unstratified sample, no items on Form S had a distractor chosen more often than a correct response.

3.1.7 Common Items

Eight items were common between Forms T1 and T2. They are displayed in Table 8. The table also displays the results of the independent samples *t*-tests comparing the *p*-values for these common items. As can be seen, all eight *t*-tests were statistically significant, but the effect sizes (Cohen's *d*) were trivial, ranging from .02 to .11. Note that the *t*-tests were run for unstratified sample only. A difference in *p*-values was also examined for the stratified samples only. The differences were small, ranging from |.01| to |.03| (see Attachment 1).

Table 8. VA Unstratified Sample Common Items Difficulty

T1 Item #	T2 Item #	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	Sig.	ES
T1-10	T2-11	.62	.49	.59	.49	.00	.06
T1-11	T2-10	.64	.48	.63	.48	.00	.03
T1-12	T2-13	.62	.49	.57	.49	.00	.10
T1-13	T2-12	.73	.44	.71	.45	.00	.06
T1-14	T2-15	.69	.46	.72	.45	.00	.06
T1-15	T2-14	.62	.49	.63	.48	.02	.02
T1-16	T2-17	.61	.49	.55	.50	.00	.11
T1-20	T2-21	.50	.50	.48	.50	.00	.03

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; Sig. = Significance Levels; ES = Effect Size.

The only technical report that included an examination of the common items was by Morath et al. (2008) for Form S. Similar to the current findings, the authors did not detect sizable differences among the common items on Form S.

3.1.8 Subgroup Differences

Table 9 summarizes the CTT effect sizes (Cohen's d) for the mean score difference among the gender, racial, and ethnic subgroups across Forms T1 and T2.

Regarding Female (F) and Male (M) subgroup comparisons, effect sizes were small on both test forms. On T1, the effect sizes ranged from .00 to .27, with M scoring higher than F on 23 items. On T2, the effect sizes ranged from .00 to .26, with M scoring higher than F on 20 items.

Regarding Black (B) and White (W) subgroup comparisons, effect sizes were small to moderate on both test forms. On T1, the effect sizes ranged from .02 to .71, with W scoring higher than B on 25 items. On T2, the effect sizes ranged from .09 to .52, with W scoring higher than B on 25 items.

Regarding Asian (A) and White (W) subgroup comparisons, effect sizes were small on both test forms. On T1, the effect sizes ranged from .01 to .34, with W scoring higher than A on 22 items. On T2, effect sizes ranged from .01 to .25, with W scoring higher than A on 22 items.

Regarding White Hispanic (WH) and White Non-Hispanic (WnH) subgroup comparisons, effect sizes were small on both test forms. On T1, the effect sizes ranged from .03 to .26, with WnH scoring higher than WH on 24 items. On T2, effect sizes ranged from .01 to .26, with WnH scoring higher than WH on 25 items.

Regarding Black Hispanic (BH) and Black Non-Hispanic (BnH) subgroup comparisons, effect sizes were small on both test forms. On T1, the effect sizes ranged from .00 to .18, with BnH scoring higher than BH on 13 items. On T2, the effect sizes ranged from .01 to .19, with BnH scoring higher than BH on 12 items.

Table 9. VA Unstratified Sample Mean Score Subgroup Differences - Cohen's *d*

Item #	T1														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.70	.79	.22**	.58	.82	.60**	.69	.82	.33**	.73	.83	.24**	.59	.59	.01
02	.71	.76	.12**	.63	.78	.36**	.69	.78	.21**	.71	.78	.18**	.63	.64	.02
03	.52	.65	.27**	.37	.68	.67**	.54	.68	.29**	.58	.69	.22**	.46	.39	.14**
04	.79	.81	.04**	.76	.81	.13**	.81	.81	.01	.78	.82	.09**	.75	.77	.04
05	.72	.75	.05**	.73	.74	.02	.76	.74	.03	.76	.74	.03 [†]	.73	.74	.02
06	.61	.64	.06**	.46	.68	.48**	.55	.68	.27**	.57	.69	.26**	.46	.47	.02
07	.64	.66	.03**	.39	.71	.71**	.59	.71	.27**	.64	.71	.17**	.50	.41	.18**
08	.72	.69	.06**	.66	.72	.14**	.57	.73	.34**	.69	.73	.09**	.62	.67	.11*
09	.57	.62	.12**	.51	.63	.26**	.58	.63	.12**	.60	.64	.09**	.57	.52	.10*
10	.60	.62	.04**	.44	.67	.49**	.57	.66	.20**	.60	.67	.15**	.50	.44	.13**
11	.62	.65	.08**	.56	.67	.23**	.63	.67	.08**	.60	.67	.15**	.55	.57	.03
12	.58	.64	.12**	.45	.66	.45**	.59	.66	.14**	.58	.67	.19**	.49	.46	.05
13	.68	.76	.18**	.64	.76	.27**	.72	.76	.08**	.71	.76	.11**	.63	.65	.04
14	.63	.71	.19**	.58	.72	.31**	.68	.71	.08**	.66	.72	.13**	.59	.58	.00
15	.59	.63	.08**	.60	.64	.08**	.62	.64	.03	.58	.64	.14**	.53	.60	.14**
16	.59	.62	.05**	.57	.62	.12**	.61	.62	.02	.59	.63	.07**	.55	.57	.03
17	.47	.51	.08**	.43	.51	.16**	.55	.51	.08**	.49	.52	.06**	.52	.43	.18**
18	.61	.64	.06**	.54	.66	.25**	.61	.65	.10**	.62	.66	.08**	.54	.53	.01
19	.53	.56	.05**	.44	.58	.29**	.51	.58	.14**	.50	.59	.17**	.41	.45	.09*
20	.53	.48	.09**	.41	.52	.23**	.50	.52	.03	.49	.52	.07**	.44	.41	.06
21	.32	.36	.09**	.25	.38	.28**	.36	.38	.05*	.30	.39	.19**	.22	.26	.08 [†]
22	.39	.41	.05**	.27	.44	.35**	.36	.44	.17**	.36	.45	.18**	.30	.28	.04
23	.37	.39	.05**	.29	.41	.26**	.36	.41	.10**	.38	.41	.07**	.33	.29	.08 [†]
24	.31	.31	.00	.23	.33	.21**	.30	.33	.06*	.28	.33	.11**	.24	.24	.00
25	.30	.31	.01	.25	.32	.15**	.33	.32	.02	.31	.32	.03 [†]	.23	.25	.04

Item #	T2														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.64	.75	.26**	.57	.77	.47**	.66	.77	.25**	.71	.77	.14**	.61	.58	.07
02	.83	.83	.00	.73	.86	.37**	.79	.86	.22**	.82	.87	.12**	.77	.74	.07
03	.74	.78	.10**	.71	.79	.17**	.77	.79	.03	.77	.78	.03 [†]	.75	.71	.08 [†]
04	.62	.68	.13**	.53	.70	.37**	.62	.70	.16**	.61	.71	.21**	.52	.54	.05
05	.67	.68	.03**	.61	.70	.20**	.68	.70	.06*	.64	.71	.15**	.59	.61	.05
06	.60	.68	.18**	.47	.71	.52**	.60	.71	.24**	.61	.72	.26**	.53	.49	.09 [†]
07	.68	.62	.11**	.56	.67	.23**	.58	.67	.19**	.57	.68	.23**	.53	.56	.06
08	.59	.61	.05**	.43	.65	.47**	.56	.65	.18**	.57	.66	.19**	.51	.44	.14**
09	.60	.59	.02 [†]	.43	.63	.43**	.57	.63	.12**	.55	.64	.19**	.46	.44	.05
10	.60	.64	.08**	.55	.65	.21**	.63	.65	.04 [†]	.59	.66	.15**	.54	.56	.04
11	.58	.60	.04**	.39	.64	.51**	.54	.64	.19**	.57	.64	.16**	.47	.40	.15**
12	.65	.73	.18**	.64	.73	.21**	.68	.73	.12**	.69	.73	.10**	.64	.63	.02
13	.53	.59	.12**	.37	.62	.51**	.54	.61	.16**	.54	.62	.17**	.48	.38	.19**
14	.60	.64	.07**	.60	.64	.09**	.62	.64	.05 [†]	.61	.65	.09**	.57	.61	.07
15	.65	.74	.20**	.60	.74	.32**	.69	.74	.12**	.69	.75	.14**	.61	.62	.01
16	.41	.50	.18**	.36	.51	.31**	.46	.51	.10**	.44	.51	.15**	.37	.36	.02
17	.55	.56	.01	.48	.57	.19**	.57	.57	.01	.56	.57	.01	.48	.49	.01
18	.62	.59	.06**	.54	.62	.17**	.59	.61	.06*	.57	.62	.10**	.50	.53	.07
19	.53	.49	.07**	.40	.53	.27**	.44	.53	.17**	.47	.53	.14**	.38	.41	.04
20	.40	.40	.01	.33	.42	.18**	.44	.41	.06*	.38	.42	.09**	.32	.33	.02
21	.51	.47	.08**	.37	.51	.28**	.51	.51	.01	.47	.51	.08**	.40	.37	.06
22	.39	.51	.24**	.37	.51	.28**	.42	.51	.18**	.45	.52	.14**	.36	.38	.05
23	.32	.33	.02 [†]	.23	.35	.26**	.35	.34	.01	.31	.35	.08**	.25	.23	.05
24	.40	.41	.03*	.33	.43	.21**	.39	.43	.07**	.38	.43	.10**	.31	.33	.06
25	.32	.37	.11**	.27	.38	.23**	.33	.38	.10**	.31	.39	.17**	.28	.27	.02

[†]*p* ≤ .100; **p* ≤ .050; ***p* ≤ .010; unmarked are non-significant *p*.

Note. *M_F* = Mean for Female Test-Takers; *M_M* = Mean for Male Test-Takers; *M_B* = Mean for Black/African American Test-Takers; *M_W* = Mean for White Test-Takers; *M_A* = Mean for Asian Test-Takers; *d* = Cohen's *d*. Underlined the effect sizes for the items for which the minority subgroups scored higher than their majority subgroups.

In general, the pattern of results was consistent among the stratified and unstratified samples, with most items showing small and some showing moderate effect sizes. On Form T1, one item had a large effect size for AECP (T1-06 for Black/White [B/W]). On Form T2, AECP had one item show a large effect size (T2-05 for WH/WnH). Similarly, AFRES had one item with large effect size (T2-05 for BH/BnH).

Table 10 displays the IRT DIF_{MH} results and Table 11 displays the Lord's chi-squared DIF method (DIF_{Lord}) results for the same subgroups. T1-01, T1-03, and T1-07 showed large effect sizes for B/W. T1-05 and T1-15 showed moderate effect sizes for B/W. T1-01 and T1-08 showed moderate effect sizes for A/W. T2-01 showed a moderate effect size for F/M. T2-01, T2-06, T2-08, T2-11, and T2-13 showed moderate to large effect size for B/W.

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 the majority White group, $N_W = 24,859$; $N_{NHPI} = 415$ compared to $N_W = 25,828$) and for Form T2 ($N_{AIAN} = 354$ compared to $N_W = 22,486$ and $N_{NHPI} = 384$ compared to $N_W = 23,481$) 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.

Table 10. VA Unstratified Sample DIF_{MH}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.93	1.52	1.45	.49	.15	1.23	1.23	.82	.00	.30
02	.30	.35	.78	.23	.08	.30	.90	.70	.21	.28
03	1.04	1.85	1.18	.28	.82	.35	.05	.31	.44	.33
04	.00	.61	.29	.08	.06	.55	.59	.28	.15	.27
05	.03	1.30	.56	.80	.22	.01	.05	.23	.09	.37
06	.01	.95	1.02	.57	.08	.88	1.39	.67	.29	.51
07	.25	1.96	1.06	.16	.79	.87	.04	.45	.39	.22
08	.65	.64	1.42	.17	.36	.11	1.18	.38	.12	.68
09	.27	.11	.30	.27	.52	.36	.95	.10	.09	.27
10	.17	.90	.70	.05	.57	.24	.07	.35	.02	.38
11	.08	.35	.04	.08	.16	.05	1.42	.43	.03	.65
12	.27	.84	.39	.33	.28	.80	.01	.11	.21	.04
13	.59	.07	.03	.07	.06	.47	1.57	.39	.23	.70
14	.71	.26	.11	.01	.02	.21	.63	.27	.22	.22
15	.08	1.12	.24	.04	.71	.89	.62	.20	.11	.10
16	.08	.65	.18	.15	.12	.80	.60	.10	.19	.21
17	.10	.59	.75	.31	.73	.09	.04	.38	.45	.01
18	.06	.29	.15	.37	.15	.47	.11	.13	.12	.36
19	.04	.11	.42	.29	.34	.50	.37	.45	.13	.18
20	.81	.36	.22	.30	.13	.08	.10	.71	.07	.15
21	.19	.11	.04	.38	.45	.57	.31	.54	.23	.24
22	.11	.31	.58	.26	.06	1.15	.35	.50	.04	.24
23	.04	.17	.26	.32	.46	.02	.49	.41	.09	.01
24	.19	.18	.03	.05	.20	.03	.07	.03	.07	.09
25	.21	.43	.36	.29	.28	.48	.22	.18	.35	.20

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Table 11. VA Unstratified Sample DIF_{Lord}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.87	1.53	1.21	.63	.01	1.06	.85	.78	.06	.18
02	.25	.33	.52	.29	.02	.41	.58	.74	.07	.15
03	.96	1.87	.91	.42	.66	.18	.61	.46	.55	.23
04	.12	.84	.62	.15	.12	.35	.33	.26	.35	.31
05	.06	1.47	.86	.80	.30	.20	.51	.30	.10	.42
06	.06	.91	.78	.65	.18	.62	1.08	.68	.58	.30
07	.28	2.03	.78	.22	.70	1.04	.38	.41	.46	.27
08	.79	.85	1.18	.20	.44	.10	.86	.37	.27	.58
09	.19	.24	.02	.27	.44	.53	.64	.06	.25	.15
10	.18	.93	.43	.06	.46	.06	.47	.39	.07	.43
11	.01	.39	.24	.12	.26	.15	1.08	.41	.11	.54
12	.17	.75	.09	.32	.20	.63	.46	.04	.16	.05
13	.53	.12	.24	.00	.17	.29	1.11	.27	.12	.68
14	.65	.03	.19	.01	.07	.02	1.14	.36	.30	.32
15	.01	1.19	.49	.05	.81	.73	.11	.06	.01	.01
16	.08	.97	.54	.30	.17	.62	.08	.08	.06	.17
17	.00	.77	1.08	.38	.70	.31	.62	.58	.67	.10
18	.11	.30	.13	.32	.06	.72	.72	.30	.26	.44
19	.14	.09	.13	.21	.42	.77	.18	.32	.00	.27
20	.91	.40	.50	.33	.05	.32	.63	.90	.19	.22
21	.10	.03	.34	.37	.49	.80	.10	.64	.23	.12
22	.15	.42	.29	.30	.04	.96	.08	.40	.02	.33
23	.12	.15	.05	.30	.39	.23	.06	.61	.24	.08
24	.33	.35	.30	.07	.17	.22	.46	.18	.17	.01
25	.39	.70	.73	.50	.33	.26	.29	.01	.27	.10

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Aguilar (2017) and Carretta et al. (2016) did not include either CTT or IRT subgroup difference analyses in their report. Morath et al. (2008) included CTT Cohen's *d* results for Form S. The authors did not observe any items with large effect sizes on S2. However, there was one item with a large effect size for B/W and Hispanic/White. There were a few items with moderate effect sizes as well.

3.1.9 Conclusion

Overall, the VA subtest items seem to be of below average difficulty and of low discriminability. Item omission rates and longstring option endorsement across the last five items suggest slight speededness and overt guessing. The removal of any one of the items would have a small effect on the internal consistency of the subtest. One item on T1 and one item on T2 had a distractor chosen more frequently than the correct response. There were some distractors that were chosen infrequently between the parallel forms. Common items analysis showed small differences in item difficulty between Forms T1 and T2. Most items showed small-to-moderate (no large) Cohen's *d* among the subgroups. Similarly, most items showed small-to-moderate DIF among the subgroups, with a few items showing large DIF for B/W subgroup comparisons.

3.2 Arithmetic Reasoning (AR)

AR is the second cognitive subtest and is one of the five subtests that make up the Officership portion of the AFOQT. It consists of 25 items with five multiple choice options. Test-takers are allotted 29 minutes to complete the test, for an average of 69.6 seconds per item.

3.2.1 Descriptives

Table 12 summarizes the means, standard deviations, skewness, and kurtosis for the unstratified sample across Forms T1 and T2. The item means ranged from .26 to .75 with an average of .53 on T1 and from .28 to .72 with an average of .53 on T2. The means generally decreased throughout the test. The means for the first five items exceeded .60, whereas the means for the last five items fell below .40 across Forms T1 and T2. Skewness tended to switch from negative to positive throughout the test, with the values remaining within the acceptable range.

Table 12. AR Unstratified Sample Descriptives

Item #	T1				T2			
	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
01	.75	.44	-1.13	-.72	.72	.45	-.96	-1.08
02	.67	.47	-.71	-1.50	.64	.48	-.60	-1.64
03	.73	.45	-1.02	-.95	.69	.46	-.82	-1.33
04	.68	.47	-.77	-1.41	.72	.45	-1.00	-1.01
05	.68	.47	-.76	-1.42	.66	.47	-.69	-1.52
06	.64	.48	-.57	-1.68	.69	.46	-.81	-1.34
07	.63	.48	-.53	-1.72	.52	.50	-.07	-2.00
08	.55	.50	-.20	-1.96	.69	.46	-.82	-1.33
09	.70	.46	-.89	-1.21	.65	.48	-.65	-1.58
10	.61	.49	-.47	-1.78	.64	.48	-.57	-1.67
11	.62	.49	-.49	-1.76	.55	.50	-.22	-1.95
12	.58	.49	-.34	-1.88	.52	.50	-.08	-1.99
13	.58	.49	-.31	-1.90	.53	.50	-.12	-1.99
14	.48	.50	.06	-2.00	.63	.48	-.55	-1.69
15	.60	.49	-.40	-1.84	.48	.50	.08	-1.99
16	.55	.50	-.19	-1.96	.45	.50	.19	-1.96
17	.48	.50	.10	-1.99	.55	.50	-.21	-1.96
18	.39	.49	.45	-1.80	.48	.50	.08	-1.99
19	.54	.50	-.16	-1.97	.42	.49	.34	-1.89
20	.41	.49	.36	-1.87	.44	.50	.23	-1.95
21	.31	.46	.80	-1.36	.29	.46	.91	-1.17
22	.28	.45	1.00	-.99	.28	.45	.98	-1.03
23	.30	.46	.89	-1.22	.33	.47	.75	-1.44
24	.33	.47	.75	-1.44	.29	.46	.91	-1.17
25	.26	.44	1.10	-.79	.30	.46	.88	-1.22

Note. *M* = Mean; *SD* = Standard Deviation. $N_{T1} = 36,097$; $N_{T2} = 32,949$.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing means and skewness switching from negative to positive throughout the test. USAFA

cadets had the highest average mean among the accession sources, with T1_{USAFA} averaging at .64 and T2_{USAFA} averaging at .65.

Carretta et al. (2016), Aguilar (2017), and Morath et al. (2008) do not report item-level central tendency statistics.

3.2.2 Difficulty

Table 13 summarizes the *p*-values and 2PL *b* parameters for the unstratified sample across Forms T1 and T2. The *p*-values ranged widely on both Forms T1 and T2, but in general, they tended to decrease throughout the test. None of the items showed ceiling or floor effect. The 2PL *b* parameters tended to switch from negative to positive throughout the test, while staying within the acceptable range.

Table 13. AR Unstratified Sample Difficulty

Item #	T1		T2	
	<i>p</i> -values	2PL <i>b</i>	<i>p</i> -values	2PL <i>b</i>
01	.75	-1.22	.72	-1.06
02	.67	-.80	.64	-.65
03	.73	-1.60	.69	-.81
04	.68	-.99	.72	-1.12
05	.68	-.92	.66	-.74
06	.64	-.63	.69	-.97
07	.63	-.80	.52	-.10
08	.55	-.26	.69	-1.26
09	.70	-.96	.65	-.68
10	.61	-.74	.64	-.84
11	.62	-.72	.55	-.28
12	.58	-.44	.52	-.12
13	.58	-.40	.53	-.17
14	.48	.05	.63	-.56
15	.60	-.44	.48	.16
16	.55	-.23	.45	.18
17	.48	.19	.55	-.25
18	.39	.46	.48	.09
19	.54	-.21	.42	.28
20	.41	.39	.44	.26
21	.31	.81	.29	.74
22	.28	.83	.28	1.00
23	.30	1.46	.33	1.60
24	.33	1.56	.29	1.56
25	.26	1.92	.30	1.36

Note. *p*-value = Difficulty; 2PL *b* = IRT Difficulty Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing *p*-values and 2PL *b* parameters switching from negative to positive throughout the test. From the CTT perspective, USAFA cadets scored higher on average than the rest of the accession sources, with T1 averaging at .64 and T2 averaging at .65. USAFA had four items on T1 and five

items on T2 that reached ceiling effect. There were a few items across the accession sources that reached floor effect (e.g., T1-22 for OTS-AD, AECP, and AFRES; T2-21 for OTS-AD; T2-22 for OTS-AD, ANG, and AFRES). Notably, all items with ceiling effect were towards the beginning of the test and all items with floor effect were towards the end of the test across the accession sources. From the IRT perspective, USAFA was the only accession source with three items on T1 (T1-01, T1-03, and T1-04) and two items on T2 (T2-01 and T2-03) having 2PL b parameters below the acceptable level.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and Aguilar (2017) for Form T. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seems to be more difficult. The mean p -values for Forms T1 and T2 were .53 and .53, respectively. The mean p -values for S1 and S2 were .63 and .64, respectively.

3.2.3 Discriminability

Table 14 summarizes the ITC's and 2PL a parameters for the unstratified sample across Forms T1 and T2. ITC's ranged from .28 to .54 with an average of .42 for T1 and from .28 to .57 with an average of .43 on T2. Eleven items on T1 and 12 items on T2 had acceptable 2PL a parameter.

Table 14. AR Unstratified Sample Discriminability

Item #	T1		T2	
	ITC	2PL a	ITC	2PL a
01	.41	1.12	.43	1.07
02	.45	1.08	.47	1.19
03	.33	.68	.48	1.33
04	.41	.92	.42	1.04
05	.41	.97	.47	1.19
06	.46	1.12	.42	.99
07	.37	.73	.44	.91
08	.44	.93	.35	.71
09	.45	1.14	.48	1.23
10	.36	.67	.38	.76
11	.39	.79	.45	.98
12	.43	.93	.43	.88
13	.45	.98	.43	.89
14	.46	1.04	.51	1.34
15	.50	1.28	.32	.54
16	.46	1.03	.49	1.19
17	.32	.54	.48	1.14
18	.49	1.21	.42	.86
19	.44	.98	.57	1.62
20	.49	1.14	.44	.97
21	.49	1.24	.56	1.73
22	.54	1.66	.47	1.20
23	.33	.64	.28	.48
24	.28	.50	.32	.61
25	.30	.60	.34	.69

Note. ITC = Item-Total Correlation; 2PL a = IRT Discriminability Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample, such that there was a wide range of ITC's among the items and the majority of the items had low 2PL a parameters. The only exception was OTS-CIV on T2, for which the number of acceptable 2PL a parameters was higher than the number of unacceptable 2PL a parameters.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and Aguilar (2017) for Form T. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seemed to be slightly more discriminable ($T1 = .41$ and $T2 = .43$ vs. $S1 = .40$ and $S2 = .41$).

3.2.4 Internal Consistency if Item is Deleted

Table 15 summarizes Cronbach's alphas for the AR subtest if item is deleted for the unstratified sample across Forms T1 and T2. For reference, the AR subtest-level alphas for T1 and for T2 were .80 and .81, respectively. The item-level alphas were high, ranging from .79 to .80 on T1 and from .81 to .82 on T2. The removal of any one of the items would have a small effect on the internal consistency of the subtest.

Table 15. AR Unstratified Sample Internal Consistency if Item is Deleted

Item #	T1 α	T2 α
01	.80	.81
02	.80	.81
03	.80	.81
04	.80	.81
05	.80	.81
06	.80	.81
07	.80	.81
08	.80	.82
09	.80	.81
10	.80	.82
11	.80	.81
12	.80	.81
13	.80	.81
14	.80	.81
15	.79	.82
16	.80	.81
17	.80	.81
18	.79	.81
19	.80	.81
20	.79	.81
21	.79	.81
22	.79	.81
23	.80	.82
24	.80	.82
25	.80	.82

Note. α = Cronbach's Alpha.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the alphas were all within the acceptable range and no one item was indicated as dramatically improving internal consistency if deleted. OTS-CIV and ROTC were the two accession sources with slightly higher alphas (in the .80 range) compared to the rest of the accession sources (in the .70 range) across Forms T1 and T2.

Although the reports by Aguilar (2017), Carretta et al. (2016), and Morath et al. (2008) did not report the values for alpha if items were deleted, the Cronbach's alpha they reported for the subtests were similar to the alphas observed in this research.

3.2.5 Omission Rates and Overt Guessing

Table 16 summarizes the percentages of omissions in a form of missing values for each item and for a grouping of five items for Forms T1 and T2. The longstring omission rate across the last five items was 2.96% for T1 and 3.24% for T2. In general, the omission rates increased throughout the test. For T1, the average omission rate on the first five items was .49%, whereas the average on the last five items was 5.70%. For T2, these averages went from .68% to 5.83%. The longstring overt guessing for the last five items was 2.51% for T1 and 2.97% for T2.

Table 16. AR Unstratified Sample Omission Rates

Item #	T1		T2	
	For Each Item	Across Five Items	For Each Item	Across Five Items
01	.58%		.34%	
02	.32%		.73%	
03	.54%	.49%	.76%	.68%
04	.72%		1.16%	
05	.30%		.42%	
06	.47%		.25%	
07	1.09%		1.61%	
08	1.39%	1.01%	.49%	.86%
09	.89%		.89%	
10	1.22%		1.08%	
11	.98%		1.27%	
12	1.17%		1.42%	
13	1.53%	1.25%	.63%	1.22%
14	1.29%		1.10%	
15	1.26%		1.70%	
16	1.94%		1.52%	
17	2.33%		2.51%	
18	2.50%	2.54%	3.02%	2.86%
19	2.77%		2.71%	
20	3.14%		4.54%	
21	4.51%		4.28%	
22	4.62%		5.18%	
23	5.58%	5.70%	6.00%	5.83%
24	6.57%		6.37%	
25	7.22%		7.30%	

Note. For Each Item = Percent of Missing Values in Each Item; Across Five Items = Percent of Missing Values Across a Grouping of five Items.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the omission rates tended to increase throughout the test. The accession sources varied in their omission rates across the last five items on both Forms T1 and T2. The lowest and the highest omission rates on T1 were observed for AECF (4.47%) and OTS-CIV (7.36%), respectively. On T2, the lowest omission rate was observed for USAFA (4.06%) and the highest omission rate was observed for OTS-CIV (7.33%).

In general, the current pattern of findings is consistent with the results reported by Morath et al. (2008) for Form S and by Carretta et al. (2016) for Form T for the unstratified sample. These reports did not include the omission and guessing rates for the stratified samples. Aguilar (2017) did not report omission rates or overt guessing.

3.2.6 Distractor Analysis

When examining the response frequencies, the correct response was chosen more frequently than a distractor for all items on both Forms T1 and T2, except for T1-23, T1-24, T2-23, and T2-24. These items happen to be common between the parallel forms and also appear towards the end of the subtest. Therefore, guessing might have played a part among these items. There were no items for which the distractor was chosen less than 1% of the time.

The pattern of results for the stratified samples was similar to the unstratified sample, such that all but one accession source had at least one item for which a distractor was chosen more often than the correct response. The exception was USAFA for Forms T1 and T2, with no distractors being chosen more often than the correct response. There were some items for which at least one distractor was chosen less than 1% of the time: T2-05 for AECF and for USAFA; T2-15 for USAFA; and T1-17 for USAFA.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) for Form T. Aguilar (2017) did not report the distractor analysis. When comparing the present results against the results for Form S reported by Morath et al. (2008) for the unstratified sample, no items on Form S had a distractor chosen more often than a correct response.

3.2.7 Common Items

Eight items were common between Forms T1 and T2. They are displayed in Table 17. The table also displays the results of the independent samples *t*-tests comparing the *p*-values for these common items. As can be seen, six *t*-tests were statistically significant, but the effect sizes (Cohen's *d*) were trivial, ranging from .02 to .13. The rest of the *t*-tests were trivial as well, with Cohen's *d* ranging from 0 to .01. Note that the *t*-tests were run for unstratified sample only. A difference in *p*-values was also examined for the stratified samples only. The differences between the *p*-values were small, ranging from |.00| to |.02| (see Attachment 1).

Table 17. AR Unstratified Sample Common Items Difficulty

T1 Item #	T2 Item #	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	Sig.	ES
T1-05	T2-06	.68	.47	.69	.46	.00	.02
T1-08	T2-07	.55	.50	.52	.50	.00	.07
T1-12	T2-11	.58	.49	.55	.50	.00	.06
T1-16	T2-17	.55	.50	.55	.50	.28	.01
T1-17	T2-15	.47	.50	.48	.50	.23	.01
T1-18	T2-16	.39	.49	.45	.50	.00	.13
T1-22	T2-21	.28	.45	.29	.46	.00	.04
T1-24	T2-23	.33	.47	.33	.47	.87	.00

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; Sig. = Significance Levels; ES = Effect Size.

The only technical report that included an examination of the common items was by Morath et al. (2008) for Form S. Similar to the current findings, the authors did not detect sizable differences among the common items on Form S.

3.2.8 Subgroup Differences

Table 18 summarizes the CTT effect sizes (Cohen's *d*) for the mean score difference among the gender, racial, and ethnic subgroups across Forms T1 and T2.

Regarding Female (F) and Male (M) subgroup comparisons, all but one item on T1 (T1-14 had moderate effect size) and one item on T2 (T2-19 had moderate effect size) showed small effect sizes. On T1, the effect sizes ranged from .00 to .47, with M scoring higher than F on 24 items. On T2, the effect sizes ranged from .08 to .41, with M scoring higher than F on 25 items.

For the Black (B) and White (W) subgroup comparisons, effect sizes were small to moderate on both test forms. On T1, the effect sizes ranged from .12 to .56, with W scoring higher than B on 25 items. On T2, the effect sizes ranged from .13 to .62, with W scoring higher than B on 25 items.

Regarding Asian (A) and White (W) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .01 to .22, with W scoring higher than A on 13 items. On T2, effect sizes ranged from .00 to .15, with W scoring higher than A on 11 items.

For the White Hispanic (WH) and White Non-Hispanic (WnH) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .02 to .23, with WnH scoring higher than WH on 25 items. On T2, effect sizes ranged from .05 to .29, with WnH scoring higher than WH on 25 items.

Regarding Black Hispanic (and Black Non-Hispanic subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .00 to .09, with BnH scoring higher than BH on 11 items. On T2, the effect sizes ranged from .00 to .13, with BnH scoring higher than BH on six items.

Table 18. AR Unstratified Sample Mean Score Subgroup Differences - Cohen's *d*

Item #	T1														
	M_F	M_M	D	M_B	M_W	d	M_A	M_W	d	M_{WH}	M_{WnH}	d	M_{BH}	M_{BnH}	d
01	.64	.78	.34**	.58	.78	.48**	.74	.78	.11**	.71	.79	.19**	.60	.59	.01
02	.63	.68	.11**	.53	.70	.35**	.76	.69	.15**	.63	.70	.14**	.55	.53	.04
03	.73	.73	.00	.65	.74	.19**	.82	.74	.19**	.73	.74	.02	.66	.65	.02
04	.60	.71	.24**	.55	.71	.34**	.72	.71	.03	.65	.71	.14**	.56	.57	.01
05	.56	.72	.34**	.53	.71	.39**	.70	.71	.02	.64	.72	.17**	.54	.54	.00
06	.55	.67	.25**	.45	.68	.48**	.67	.67	.01	.61	.68	.16**	.47	.47	.02
07	.58	.64	.13**	.52	.65	.27**	.65	.65	.01	.61	.65	.09**	.55	.53	.04
08	.49	.57	.16**	.43	.58	.30**	.59	.58	.02	.52	.58	.14**	.42	.45	.06
09	.61	.74	.28**	.54	.73	.43**	.77	.73	.10**	.69	.73	.10**	.59	.55	.08 [†]
10	.57	.63	.11**	.46	.65	.39**	.60	.65	.09**	.58	.65	.15**	.47	.48	.01
11	.58	.63	.11**	.51	.64	.28**	.70	.64	.13**	.61	.64	.07**	.49	.51	.05
12	.49	.62	.27**	.44	.61	.37**	.64	.61	.07**	.55	.62	.13**	.42	.45	.05
13	.48	.61	.26**	.42	.62	.42**	.55	.62	.14**	.54	.62	.17**	.39	.43	.09*
14	.31	.55	.47**	.30	.53	.47**	.42	.53	.22**	.43	.54	.22**	.32	.32	.00
15	.49	.64	.31**	.38	.65	.56**	.63	.64	.02	.55	.65	.21**	.40	.40	.01
16	.45	.58	.25**	.40	.59	.38**	.57	.58	.03	.50	.59	.18**	.43	.40	.07
17	.43	.49	.13**	.35	.50	.31**	.49	.50	.01	.44	.51	.13**	.38	.36	.03
18	.29	.43	.29**	.19	.43	.50**	.45	.43	.04 [†]	.35	.44	.18**	.22	.19	.08 [†]
19	.48	.56	.17**	.39	.57	.38**	.59	.57	.03	.48	.58	.22**	.38	.40	.04
20	.34	.44	.21**	.27	.44	.34**	.49	.43	.12**	.38	.44	.13**	.28	.27	.02
21	.20	.36	.35**	.15	.35	.43**	.32	.35	.06**	.25	.36	.23**	.18	.16	.07
22	.18	.31	.29**	.11	.31	.45**	.36	.31	.11**	.21	.32	.22**	.14	.11	.08 [†]
23	.23	.32	.20**	.21	.32	.24**	.30	.32	.04	.26	.32	.13**	.19	.21	.05
24	.28	.34	.14**	.26	.34	.18**	.28	.34	.12**	.29	.35	.12**	.23	.27	.07
25	.26	.26	.01	.22	.27	.12**	.31	.27	.09**	.24	.27	.06**	.18	.21	.07

Item #	T2														
	M_F	M_M	d	M_B	M_W	d	M_A	M_W	d	M_{WH}	M_{WnH}	d	M_{BH}	M_{BnH}	d
1	.62	.75	.29**	.56	.75	.42**	.74	.75	.01	.70	.75	.12**	.59	.58	.02
2	.54	.68	.28**	.44	.69	.52**	.68	.68	.00	.61	.69	.17**	.48	.46	.04
3	.60	.72	.25**	.50	.74	.53**	.71	.73	.04	.64	.74	.23**	.50	.51	.03
4	.66	.75	.20**	.59	.75	.38**	.75	.75	.01	.69	.76	.15**	.63	.59	.08
5	.57	.70	.28**	.51	.69	.37**	.75	.68	.14**	.64	.69	.12**	.56	.53	.08
6	.57	.73	.34**	.53	.72	.44**	.73	.72	.01	.66	.73	.13**	.59	.53	.12*
7	.47	.53	.13**	.40	.54	.29**	.54	.54	.01	.49	.55	.13**	.41	.41	.01
8	.66	.70	.08**	.59	.71	.25**	.74	.70	.07**	.69	.71	.05*	.62	.59	.06
9	.58	.68	.22**	.52	.68	.33**	.74	.68	.15**	.65	.68	.07**	.52	.53	.01
10	.59	.65	.13**	.52	.67	.30**	.66	.66	.01	.60	.67	.15**	.56	.53	.07
11	.46	.59	.27**	.39	.58	.38**	.64	.58	.12**	.53	.58	.11**	.42	.41	.02
12	.44	.55	.22**	.38	.55	.32**	.58	.54	.07**	.51	.54	.08**	.41	.39	.05
13	.43	.56	.26**	.36	.57	.42**	.55	.57	.03	.48	.57	.18**	.43	.37	.11*
14	.54	.67	.27**	.44	.68	.50**	.65	.67	.05*	.59	.68	.19**	.50	.46	.08
15	.43	.50	.13**	.36	.51	.30**	.47	.51	.06**	.46	.51	.10**	.36	.38	.04
16	.36	.49	.26**	.25	.50	.50**	.50	.49	.01	.41	.50	.18**	.32	.26	.13**
17	.45	.59	.27**	.39	.59	.40**	.58	.59	.01	.50	.59	.18**	.40	.39	.02
18	.41	.50	.19**	.38	.50	.25**	.57	.50	.14**	.44	.50	.12**	.36	.38	.03
19	.27	.47	.41**	.17	.47	.62**	.42	.47	.10**	.34	.48	.29**	.19	.19	.00
20	.38	.47	.16**	.33	.47	.28**	.52	.46	.12**	.42	.47	.10**	.36	.33	.06
21	.19	.33	.30**	.13	.33	.45**	.37	.32	.10**	.23	.33	.22**	.15	.14	.04
22	.21	.30	.20**	.15	.31	.35**	.35	.30	.10**	.23	.31	.18**	.17	.16	.03
23	.29	.34	.11**	.28	.34	.13**	.30	.34	.09**	.30	.34	.10**	.28	.27	.02
24	.22	.32	.21**	.22	.31	.20**	.33	.31	.04 [†]	.24	.32	.16**	.26	.22	.09 [†]
25	.23	.32	.20**	.19	.32	.29**	.32	.32	.01	.25	.32	.15**	.19	.20	.02

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

Note. M_F = Mean for Female Test-Takers; M_M = Mean for Male Test-Takers; M_B = Mean for Black/African American Test-Takers; M_W = Mean for White Test-Takers; M_A = Mean for Asian Test-Takers; d = Cohen's d . Underlined the effect sizes for the items for which the minority subgroups scored higher than their majority subgroups.

The general pattern of results was consistent among the stratified and unstratified samples, such that ethnic and A/W subgroups revealed no large or moderate effect sizes across all items. There were a few exceptions to that. Specifically, AECP had several items on Forms T1 and T2 with moderate effect sizes for A/W and WH/WnH subgroups. One item on T2 had a large effect size (T2-03 Cohen's $d = .80$). USAFA, ANG, and AFRES showed a few more items with moderate effect sizes on T2, but not on T1. Similar to the unstratified sample, Asian test-takers outscored their White counterparts on many items across the stratifications on Forms T1 and T2. For F/M subgroups, the most number of items with moderate effect sizes was eight for AECP on T2. The rest of the accession sources had either no or up to two items with moderate effect sizes and none with large effect sizes. For B/W subgroups, the number of items with moderate and large effect sizes was higher than for any other subgroups. The number of moderate effect sizes ranged from two items on T1 for ANG and on T2 for OTS-AD to 13 on T1 for AECP.

Table 19 displays the IRT DIF_{MH} results and Table 20 displays the DIF_{Lord} results for the same subgroups. T1-03 and T1-14 showed moderate effect sizes for both F/M and for A/W. T1-25 showed moderate effect size for F/M only. T2-19 showed moderate to large effect sizes for B/W.

Table 19. AR Unstratified Sample DIF_{MH}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.51	.55	.58	.15	.14	.11	.03	.06	.23	.05
02	.93	.19	1.02	.16	.09	.06	.40	.11	.03	.13
03	1.26	.67	1.28	.51	.21	.28	.36	.08	.27	.35
04	.05	.14	.24	.09	.17	.32	.16	.20	.03	.18
05	.48	.16	.10	.12	.01	.07	.40	1.08	.30	.02
06	.15	.47	.06	.05	.07	.43	.16	.18	.10	.22
07	.58	.32	.07	.26	.21	.69	.58	.03	.15	.22
08	.56	.35	.11	.10	.32	.90	.51	.51	.35	.10
09	.09	.24	.71	.30	.41	.42	.62	1.10	.63	.42
10	.64	.37	.55	.09	.08	.56	.35	.01	.02	.23
11	.68	.35	.75	.37	.14	.03	.08	.76	.21	.05
12	.07	.02	.40	.12	.08	.25	.39	.40	.38	.35
13	.04	.27	.85	.05	.38	.01	.13	.17	.07	.29
14	1.14	.50	1.39	.34	.18	.20	.16	.22	.07	.06
15	.10	.80	.13	.18	.08	.47	.21	.36	.05	.24
16	.13	.00	.26	.14	.19	.21	.54	.04	.06	.45
17	.43	.10	.21	.06	.09	.03	.12	.06	.10	.16
18	.01	.92	.10	.14	.36	.41	.70	.77	.21	.38
19	.57	.14	.14	.29	.21	.63	1.38	.67	.61	.37
20	.44	.23	.55	.22	.18	.62	.65	.63	.33	.18
21	.54	.64	.71	.52	.39	.15	.41	.55	.29	.12
22	.06	.75	.44	.34	.47	.49	.15	.48	.21	.10
23	.09	.12	.43	.03	.31	.47	.77	.62	.08	.08
24	.28	.46	.88	.01	.35	.05	.53	.10	.29	.22
25	1.11	.84	.28	.25	.56	.11	.01	.09	.15	.10

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Table 20. AR Unstratified Sample DIF_{Lord}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.74	.59	.64	.24	.08	.39	.22	.18	.13	.08
02	.59	.12	.93	.07	.03	.29	.70	.13	.09	.10
03	1.22	.90	1.30	.67	.14	.14	.74	.32	.42	.30
04	.13	.19	.17	.06	.12	.07	.00	.03	.09	.17
05	.67	.07	.14	.14	.06	.28	.07	.79	.18	.01
06	.16	.52	.09	.03	.10	.65	.27	.05	.05	.21
07	.50	.56	.09	.32	.21	.56	.50	.21	.12	.19
08	.33	.44	.09	.05	.34	.82	.66	.36	.47	.07
09	.39	.28	.62	.19	.36	.06	.27	.84	.45	.34
10	.62	.05	.57	.02	.07	.56	.46	.18	.01	.21
11	.61	.55	.76	.45	.16	.19	.01	.58	.19	.09
12	.25	.08	.38	.10	.09	.10	.34	.23	.40	.38
13	.18	.18	.81	.10	.41	.15	.19	.35	.06	.33
14	1.40	.56	1.36	.39	.16	.20	.56	.41	.25	.04
15	.48	.95	.14	.30	.05	.60	.48	.55	.24	.27
16	.16	.01	.23	.17	.23	.15	.79	.14	.16	.44
17	.54	.37	.15	.15	.13	.22	.09	.24	.19	.15
18	.42	1.07	.14	.20	.35	.26	.74	.64	.21	.41
19	.32	.02	.14	.31	.18	1.06	1.82	.76	.78	.25
20	.09	.10	.59	.11	.21	.41	.54	.48	.30	.16
21	.92	.80	.50	.60	.39	.63	1.05	.34	.50	.05
22	.65	1.25	.52	.54	.45	.11	.29	.34	.31	.09
23	.00	.61	.37	.13	.29	.68	1.40	.84	.28	.15
24	.47	1.04	.91	.23	.34	.04	.94	.01	.16	.20
25	1.27	1.40	.38	.49	.57	.12	.32	.19	.03	.17

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Aguilar (2017) and Carretta et al. (2016) did not include either CTT or IRT subgroup difference analyses in their report. Morath et al. (2008) included CTT Cohen's *d* results for Form S. The authors did not observe any items with large effect sizes. However, there were items with moderate subgroup differences, with the highest number of such items observed for B/W.

3.2.9 Conclusion

Overall, the items in the AR subtest seem to be of average difficulty and of low-to-moderate discriminability. Item omission rates and longstring option endorsement across the last five items suggest slight speededness and overt guessing. The removal of any one of the items would have a small effect on the internal consistency of the subtest. There were two common items between Forms T1 and T2 with distractors chosen more frequently than the correct response. No distractors were chosen infrequently. Common items analysis showed small differences in item difficulty between Forms T1 and T2. Most items showed small-to-moderate (no large) Cohen's *d* among the subgroups. Similarly, most items showed small-to-moderate DIF among the subgroups, with one item on T2 showing large DIF for B/W subgroups.

3.3 Word Knowledge (WK)

WK is the third cognitive subtest presented to test-takers and is one of the five subtests that make up the Officership portion of the AFOQT. It consists of 25 items with five multiple choice options. Test-takers are allotted 5 minutes to complete the test, for an average of 12 seconds per item.

3.3.1 Descriptives

Table 21 summarizes the means, standard deviations, skewness, and kurtosis for the unstratified sample across Forms T1 and T2. The item means ranged from .31 to .83 with an average of .59 on T1 and from .28 to .84 with an average of .58 on T2. The means generally decreased throughout the test. For both Forms T1 and T2, the means for the first five items exceeded .60, whereas the means for the last five items fell below .60. Skewness tended to switch from negative to positive throughout the test, with the values remaining within the acceptable range.

Table 21. WK Unstratified Sample Descriptives

Item #	T1				T2			
	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
01	.76	.43	-1.23	-.48	.84	.37	-1.85	1.43
02	.83	.38	-1.76	1.09	.70	.46	-.87	-1.24
03	.77	.42	-1.29	-.34	.78	.41	-1.35	-.17
04	.76	.43	-1.23	-.49	.68	.47	-.76	-1.43
05	.62	.49	-.49	-1.76	.69	.46	-.82	-1.33
06	.68	.47	-.78	-1.39	.68	.47	-.76	-1.42
07	.65	.48	-.63	-1.60	.63	.48	-.55	-1.70
08	.59	.49	-.38	-1.85	.68	.47	-.77	-1.40
09	.70	.46	-.87	-1.24	.58	.49	-.30	-1.91
10	.65	.48	-.64	-1.58	.70	.46	-.88	-1.22
11	.58	.49	-.34	-1.89	.63	.48	-.55	-1.70
12	.63	.48	-.55	-1.70	.62	.49	-.48	-1.77
13	.57	.50	-.28	-1.92	.64	.48	-.60	-1.64
14	.57	.49	-.30	-1.91	.56	.50	-.24	-1.94
15	.58	.49	-.33	-1.89	.54	.50	-.17	-1.97
16	.54	.50	-.16	-1.97	.56	.50	-.23	-1.95
17	.53	.50	-.13	-1.98	.52	.50	-.08	-1.99
18	.41	.49	.35	-1.88	.59	.49	-.37	-1.87
19	.53	.50	-.10	-1.99	.52	.50	-.08	-1.99
20	.64	.48	-.57	-1.68	.46	.50	.15	-1.98
21	.43	.49	.30	-1.91	.35	.48	.62	-1.61
22	.56	.50	-.25	-1.94	.47	.50	.11	-1.99
23	.52	.50	-.07	-1.99	.37	.48	.54	-1.71
24	.43	.49	.29	-1.92	.28	.45	1.00	-1.01
25	.31	.46	.85	-1.28	.45	.50	.21	-1.96

Note. *M* = Mean; *SD* = Standard Deviation. $N_{T1} = 36,094$; $N_{T2} = 32,945$.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing means and skewness switching from negative to positive throughout the test. USAFA cadets had the highest average mean than the rest of the accession sources, with T1 averaging at

.65 and T2 averaging at .63. Carretta et al. (2016), Aguilar (2017), and Morath et al. (2008) do not report item-level central tendency statistics.

3.3.2 Difficulty

Table 22 summarizes the p -values and 2PL b parameters for the unstratified sample across Forms T1 and T2. The p -values ranged widely on both Forms T1 and T2, but in general, they tended to decrease throughout the test. None of the items showed a floor effect. One item on T1 (T1-02) and one item on T2 (T2-01) showed a ceiling effect. The 2PL b parameters tended to switch from negative to positive throughout the test, while staying within the acceptable range.

Table 22. WK Unstratified Sample Difficulty

Item #	T1		T2	
	p -values	2PL b	p -values	2PL b
01	.76	-1.11	.84	-1.38
02	.83	-1.98	.70	-1.28
03	.77	-1.41	.78	-1.84
04	.76	-1.17	.68	-.77
05	.62	-.45	.69	-.84
06	.68	-1.05	.68	-.74
07	.65	-.72	.63	-.54
08	.59	-.48	.68	-.89
09	.70	-.76	.58	-.39
10	.65	-.57	.70	-.74
11	.58	-.37	.63	-.50
12	.63	-.79	.62	-.50
13	.57	-.36	.64	-.81
14	.57	-.30	.56	-.32
15	.58	-.46	.54	-.21
16	.54	-.20	.56	-.29
17	.53	-.18	.52	-.12
18	.41	.49	.59	-.58
19	.53	-.14	.52	-.11
20	.64	-.54	.46	.10
21	.43	.46	.35	.77
22	.56	-.28	.47	.14
23	.52	-.11	.37	.55
24	.43	.25	.28	1.04
25	.31	.84	.45	.48

Note. p -value = Difficulty; 2PL b = IRT Difficulty Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing p -values and 2PL b parameters switching from negative to positive throughout the test. From the CTT perspective, USAFA cadets scored higher on average than the rest of the accession sources, with T1 averaging at .65 and T2 averaging at .63. All accession sources had at least one item with a ceiling effect. No items across the accession sources reached a floor effect. Notably, all items with a ceiling effect were towards the beginning of the test. From the IRT perspective,

AECP, USAFA, and ROTC had one item on T1 with too low 2PL b parameter. OTS-AD and USAFA had at least one item with too low 2PL b parameter.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and Aguilar (2017) for Form T. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seems to be more difficult. The mean p -values for Forms T1 and T2 were .59 and .58, respectively. The mean p -values for S1 and S2 were .63 and .62, respectively.

3.3.3 Discriminability

Table 23 summarizes the ITC's and 2PL a parameters for the unstratified sample across Forms T1 and T2. ITC's ranged from .33 to .60 with an average of .47 for T1 and from .28 to .60 with an average of .46 on T2. Thirteen items on T1 and 15 items on T2 had acceptable 2PL a parameter.

Table 23. WK Unstratified Sample Discriminability

Item #	T1		T2	
	ITC	2PL a	ITC	2PL a
01	.48	1.45	.46	1.79
02	.33	.93	.36	.74
03	.40	1.05	.33	.77
04	.46	1.31	.49	1.28
05	.60	1.87	.48	1.26
06	.39	.85	.51	1.42
07	.46	1.09	.54	1.47
08	.45	.98	.44	1.06
09	.55	1.78	.44	.97
10	.59	1.87	.57	1.91
11	.52	1.33	.59	1.81
12	.39	.79	.52	1.33
13	.46	1.00	.40	.86
14	.59	1.74	.44	.94
15	.40	.83	.47	1.07
16	.49	1.13	.46	1.04
17	.42	.87	.44	.93
18	.41	.79	.37	.71
19	.47	1.04	.48	1.10
20	.55	1.55	.60	1.74
21	.36	.69	.42	.93
22	.50	1.21	.42	.86
23	.45	.94	.50	1.21
24	.55	1.45	.46	1.15
25	.49	1.26	.28	.45

Note. ITC = Item-Total Correlation; 2PL a = IRT Discriminability Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample, such that there was a wide range of ITC's among the items and the majority of the items had high 2PL a parameters. The number of items with high 2PL a parameters varied among the accession sources across Forms T1 and T2, ranging from 13 to 20 on T1 and from 13 to 18 on T2.

In general, the current pattern is consistent with the results reported by Carretta et al. (2016) and Aguilar (2017) for Form T. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seemed to be slightly more discriminable ($T1 = .47$ and $T2 = .46$ vs. $S1 = .40$ and $S2 = .42$).

3.3.4 Internal Consistency if Item is Deleted

Table 24 summarizes Cronbach's alphas if an item is deleted for the unstratified sample across Forms T1 and T2. For reference, the WK subtest-level alphas for T1 and for T2 were .85 and .84, respectively. The item-level alphas were high, ranging from .84 to .85 on T1 and from .83 to .85 on T2. The removal of any one of the items would have a small effect on the internal consistency of the subtest.

Table 24. WK Unstratified Sample Internal Consistency if Item is Deleted

Item #	T1 α	T2 α
01	.85	.84
02	.85	.84
03	.85	.84
04	.85	.84
05	.84	.84
06	.85	.84
07	.85	.84
08	.85	.84
09	.85	.84
10	.84	.83
11	.85	.83
12	.85	.84
13	.85	.84
14	.84	.84
15	.85	.84
16	.85	.84
17	.85	.84
18	.85	.84
19	.85	.84
20	.85	.83
21	.85	.84
22	.85	.84
23	.85	.84
24	.85	.84
25	.85	.85

Note. α = Cronbach's Alpha.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the alphas were all within the acceptable range and no one item was indicated as dramatically improving internal consistency if deleted.

Although the reports by Aguilar (2017), Carretta et al. (2016), and Morath et al. (2008) did not report the values for alpha if items were deleted, the Cronbach's alpha they reported for the subtests were similar to the alphas observed in this research.

3.3.5 Omission Rates and Overt Guessing

Table 25 summarizes the percentages of the omission rates in a form of missing values for each item and for a grouping of five items for Forms T1 and T2. The longstring omission rate across the last five items was 1.05% for T1 and .95% for T2. In general, the omission rates increased throughout the test, going from an average of .41% on the first five items to 2.55% on the last five items for T1, and from .66% to 2.70% on T2. The longstring overt guessing for the last five items was .91% for T1 and 1.20% for T2.

Table 25. WK Unstratified Sample Omission Rates

Item #	T1		T2	
	For Each Item	Across Five Items	For Each Item	Across Five Items
01	.62%	.41%	.66%	.66%
02	.22%		.44%	
03	.40%		.59%	
04	.35%		.93%	
05	.46%		.69%	
06	.42%	.47%	.85%	.49%
07	.37%		.22%	
08	.66%		.18%	
09	.39%		.85%	
10	.48%		.37%	
11	.68%	.59%	.82%	.73%
12	.30%		.58%	
13	.61%		.48%	
14	.50%		.65%	
15	.85%		1.14%	
16	1.12%	1.27%	.68%	1.24%
17	.77%		1.68%	
18	1.84%		.87%	
19	1.03%		1.23%	
20	1.57%		1.72%	
21	2.18%	2.55%	1.38%	2.7%
22	1.76%		1.94%	
23	2.53%		3.24%	
24	3.21%		2.97%	
25	3.06%		3.95%	

Note. For Each Item = Percent of Missing Values in Each Item; Across Five Items = Percent of Missing Values Across a Grouping of five Items.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the omission rates tended to increase throughout the test. The accession sources varied in their omission rates across the last five items on both Forms T1 and T2. The lowest and the highest omission rates on T1 were observed for USAFA (.93%) and for AECP (3.88%), respectively. On T2, the lowest omission rate was observed for USAFA (1.17%) and the highest omission rate was observed for AECP (3.92%).

In general, the current pattern of findings is consistent with the results reported by Morath et al. (2008) for Form S and Carretta et al. (2016) for Form T for the unstratified sample. These reports did not include the omission and guessing rates for the stratified samples. Aguilar (2017) did not report omission rates or overt guessing.

3.3.6 Distractor Analysis

When examining the response frequencies, the correct response was chosen more frequently than a distractor for all items on T1 and for all but one item on T2 (T2-24). The only item for which a distractor was chosen less than 1% of the time was T1-02.

The pattern of results for the stratified samples was similar to the unstratified sample. Specifically, all but two accession sources (AECP and ROTC) selected the correct response most frequently on all T1 items. All but three accession sources (OTS-AD, AECP, and ROTC) selected the correct response most frequently on all T2 items. There were some items for which at least one distractor was chosen less than 1% of the time (e.g., T1-02 for OTS-CIV; T2-03 for AECP).

In general, the current findings are consistent with the results reported by Carretta et al. (2016) for Form T. Aguilar (2017) did not report the distractor analysis. When comparing the present results against those for Form S reported by Morath et al. (2008) for the unstratified sample, no items on Form S had a distractor chosen more often than a correct response.

3.3.7 Common Items

Nine items were common between Forms T1 and T2. They are displayed in Table 26. The table also displays the results of the independent samples *t*-tests comparing the *p*-values for these common items. As can be seen, seven *t*-tests were statistically significant, but the effect sizes (Cohen's *d*) were trivial, ranging from .02 to .22. The rest of the *t*-tests were trivial, with Cohen's *d* being equal to 0. Note that the *t*-tests were run for unstratified sample only. A difference in *p*-values was also examined for the stratified samples only. The differences between the *p*-values were small, ranging from |.00| to |.03| (see Attachment 1).

Table 26. WK Unstratified Sample Common Items Difficulty

T1 Item #	T2 Item #	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	Sig.	ES
T1-08	T2-09	.59	.49	.57	.49	.00	.04
T1-09	T2-10	.70	.46	.70	.46	.75	.00
T1-10	T2-11	.65	.48	.63	.48	.00	.04
T1-11	T2-12	.58	.49	.62	.49	.00	.07
T1-13	T2-14	.57	.49	.56	.50	.00	.02
T1-16	T2-15	.54	.50	.54	.50	.92	.00
T1-18	T2-17	.41	.49	.52	.50	.00	.22
T1-23	T2-22	.52	.50	.47	.50	.00	.09
T1-25	T2-24	.31	.46	.28	.45	.00	.06

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; Sig. = Significance Levels; ES = Effect Size.

The only technical report that included an examination of the common items was by Morath et al. (2008) for Form S. Similar to the current findings, the authors did not detect sizable differences among the common items on Form S.

3.3.8 Subgroup Differences

Table 27 summarizes the CTT effect sizes (Cohen's d) for the mean score difference among the gender, racial, and ethnic subgroups across Forms T1 and T2.

For the Female (F) and Male (M) subgroup comparisons, effect sizes were mostly small on both test forms. On T1, the effect sizes ranged from .01 to .43 (with T1-14 showing a moderate effect size), with M scoring higher than F on 20 items. On T2, the effect sizes ranged from .01 to .28, with M scoring higher than F on 21 items.

Regarding Black (B) and White (W) subgroup comparisons, effect sizes were small to moderate on both test forms. On T1, the effect sizes ranged from .05 to .61, with W scoring higher than B on 24 items. Eight items showed moderate effect sizes. On T2, the effect sizes ranged from .06 to .65, with W scoring higher than B on 25 items. Six items showed moderate effect sizes.

For the Asian (A) and White (W) subgroup comparisons, effect sizes were mostly small on both test forms. On T1, the effect sizes ranged from .01 to .42 (with T1-03 showing a moderate effect size), with W scoring higher than A on 23 items. On T2, the effect sizes ranged from .05 to .50 (with T2-01 showing a moderate effect size), with W scoring higher than A on 24 items.

Regarding White Hispanic (WH) and White Non-Hispanic (WnH) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .02 to .32, with WnH scoring higher than WH on 24 items. On T2, effect sizes ranged from .01 to .29, with WnH scoring higher than WH on 24 items.

For the Black Hispanic (BH) and Black Non-Hispanic (BnH) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .00 to .26, with BnH scoring higher than BH on eight items. On T2, effect sizes ranged from .01 to .26, with BnH scoring higher than BH on five items.

Table 27. WK Unstratified Sample Mean Score Subgroup Differences – Cohen's d

Item #	T1														
	M_F	M_M	d	M_B	M_W	d	M_A	M_W	d	M_{WH}	M_{WnH}	d	M_{BH}	M_{BnH}	d
01	.74	.77	.07**	.58	.80	.54**	.72	.80	.19**	.73	.81	.21**	.64	.60	.07
02	.80	.84	.10**	.76	.84	.22**	.84	.84	.01	.83	.84	.04†	.82	.77	.12**
03	.71	.79	.19**	.58	.82	.61**	.66	.82	.42**	.71	.84	.32**	.60	.61	.00
04	.76	.76	.01	.83	.76	.15**	.73	.76	.07**	.72	.77	.13**	.74	.83	.23**
05	.59	.63	.09**	.40	.67	.58**	.56	.67	.23**	.59	.68	.19**	.46	.42	.08†
06	.64	.70	.12**	.62	.71	.21**	.54	.72	.39**	.63	.72	.20**	.57	.63	.14**
07	.64	.66	.03**	.50	.69	.40**	.58	.69	.24**	.60	.70	.21**	.52	.52	.00
08	.60	.59	.01	.49	.63	.29**	.49	.63	.30**	.56	.64	.15**	.47	.51	.09†
09	.63	.73	.21**	.56	.74	.40**	.61	.74	.30**	.67	.75	.17**	.57	.57	.01
10	.59	.68	.18**	.49	.70	.45**	.59	.70	.22**	.62	.70	.19**	.55	.51	.08†
11	.53	.60	.14**	.39	.64	.51**	.50	.63	.27**	.53	.64	.23**	.44	.40	.07
12	.70	.61	.18**	.55	.67	.26**	.51	.67	.34**	.59	.68	.19**	.54	.56	.03
13	.50	.59	.19**	.42	.60	.37**	.53	.60	.15**	.55	.61	.12**	.48	.44	.09*
14	.42	.63	.43**	.37	.62	.51**	.53	.62	.18**	.56	.62	.13**	.49	.39	.21**
15	.52	.60	.16**	.52	.60	.15**	.60	.60	.02	.54	.60	.13**	.53	.53	.00
16	.52	.55	.06**	.39	.58	.37**	.50	.58	.15**	.49	.59	.19**	.43	.40	.06
17	.50	.54	.10**	.43	.55	.23**	.53	.55	.03	.56	.55	.03†	.57	.44	.26**
18	.42	.41	.01	.29	.45	.32**	.38	.45	.13**	.40	.45	.10**	.36	.29	.16**
19	.61	.49	.24**	.51	.54	.05**	.53	.53	.01	.50	.54	.08**	.46	.51	.10*
20	.60	.65	.11**	.50	.67	.36**	.63	.67	.08**	.62	.67	.11**	.56	.51	.10*
21	.44	.42	.05**	.31	.46	.30**	.39	.45	.13**	.39	.46	.14**	.33	.32	.03
22	.52	.58	.11**	.41	.59	.37**	.62	.58	.07**	.58	.59	.02	.50	.41	.19**
23	.47	.54	.14**	.41	.55	.28**	.44	.55	.22**	.47	.56	.18**	.42	.43	.02
24	.39	.44	.10**	.30	.46	.32**	.43	.46	.05*	.38	.47	.17**	.33	.32	.02
25	.25	.33	.17**	.16	.34	.38**	.25	.34	.19**	.27	.35	.15**	.21	.18	.07
Item #	T2														
	M_F	M_M	d	M_B	M_W	d	M_A	M_W	d	M_{WH}	M_{WnH}	d	M_{BH}	M_{BnH}	d
01	.79	.86	.20**	.69	.89	.61**	.72	.89	.50**	.80	.90	.29**	.73	.70	.08†
02	.66	.72	.13**	.58	.74	.34**	.61	.74	.28**	.67	.74	.16**	.58	.59	.02
03	.73	.80	.16**	.69	.80	.28**	.76	.80	.10**	.79	.80	.03	.75	.69	.12**
04	.61	.70	.18**	.54	.71	.36**	.68	.70	.05*	.64	.71	.16**	.56	.54	.02
05	.64	.71	.15**	.59	.72	.29**	.65	.72	.15**	.67	.72	.11**	.60	.59	.00
06	.66	.69	.06**	.44	.73	.65**	.65	.73	.18**	.65	.74	.21**	.53	.47	.12*
07	.62	.64	.04**	.53	.67	.30**	.56	.67	.24**	.57	.68	.23**	.58	.54	.09†
08	.62	.70	.19**	.68	.70	.06**	.56	.70	.31**	.62	.71	.21**	.57	.69	.26**
09	.59	.57	.03*	.48	.62	.27**	.44	.61	.36**	.55	.62	.16**	.44	.49	.11*
10	.62	.73	.25**	.55	.75	.44**	.59	.75	.35**	.67	.75	.20**	.58	.55	.05
11	.56	.66	.20**	.46	.68	.46**	.58	.68	.21**	.58	.69	.23**	.50	.48	.04
12	.56	.64	.15**	.40	.67	.58**	.53	.67	.29**	.58	.68	.22**	.44	.42	.04
13	.61	.65	.09**	.50	.67	.36**	.60	.67	.15**	.64	.68	.07**	.56	.51	.10*
14	.48	.59	.21**	.42	.60	.35**	.51	.59	.17**	.55	.60	.09**	.50	.43	.15**
15	.52	.55	.05**	.39	.58	.38**	.50	.58	.16**	.51	.59	.16**	.43	.41	.03
16	.46	.59	.28**	.44	.58	.29**	.56	.58	.05*	.59	.57	.03	.54	.44	.22**
17	.53	.52	.01	.37	.57	.41**	.46	.56	.20**	.49	.57	.16**	.38	.37	.00
18	.60	.59	.03*	.46	.62	.34**	.58	.62	.08**	.57	.63	.12**	.48	.46	.04
19	.46	.54	.17**	.45	.55	.19**	.46	.54	.17**	.48	.55	.15**	.41	.46	.09*
20	.41	.48	.16**	.31	.50	.39**	.47	.50	.05*	.42	.51	.17**	.34	.32	.04
21	.27	.38	.23**	.26	.37	.24**	.34	.37	.07**	.36	.37	.03	.28	.26	.04
22	.41	.49	.16**	.38	.50	.24**	.41	.50	.17**	.44	.51	.13**	.39	.39	.01
23	.28	.40	.27**	.25	.40	.31**	.34	.40	.12**	.39	.40	.01	.36	.25	.26**
24	.23	.30	.15**	.17	.31	.30**	.23	.31	.18**	.24	.31	.16**	.18	.18	.02
25	.46	.44	.04**	.36	.47	.23**	.44	.47	.07**	.44	.47	.07**	.33	.37	.08

† $p \leq .100$; * $p \leq .050$; ** $p \leq .01$; unmarked are non-significant p .

Note. M_F = Mean for Female Test-Takers; M_M = Mean for Male Test-Takers; M_B = Mean for Black/African American Test-Takers; M_W = Mean for White Test-Takers; M_A = Mean for Asian Test-Takers; d = Cohen's d . Underlined the effect sizes for the items for which the minority subgroups scored higher than their majority subgroups.

The general pattern of results was consistent among the stratified and unstratified samples, with most items showing small and some showing moderate effect sizes. On T1, the following items had large effect sizes: OTS-AD for A/W T1-03; AECF for A/W T1-03, T1-06, and T1-13; and AECF for B/W T1-18. On T2, the following items had large effect sizes: OTS-AD for A/W T2-01; AECF for B/W T2-01 and T2-12; and AECF for WH/WnH T2-17 and T2-20.

Table 28 displays the IRT DIF_{MH} results and Table 29 displays the DIF_{Lord} results for the same subgroups. On T1, consistently moderate-to-large effect sizes were observed for three items for F/M, for five items for B/W, for three items for A/W, for one item for WH/WnH, and for one item for BH/BnH. On T2, consistently moderate-to-large DIF were observed for three items for B/W, for one item for A/W, and for two items for BH/BnH.

Table 28. WK Unstratified Sample DIF_{MH}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.14	1.20	.14	.34	.36	.24	1.08	1.94	.75	.03
02	.25	.13	.84	.46	.49	.05	.20	.75	.17	.31
03	.75	1.68	1.46	1.06	.23	.23	.02	.15	.47	.29
04	.38	2.84	.71	.24	1.20	.04	.21	.75	.11	.07
05	.12	1.30	.17	.06	.29	.14	.52	.17	.38	.25
06	.20	.27	1.22	.38	.81	.77	1.30	.02	.17	.35
07	.30	.53	.35	.33	.11	.95	.65	.32	.27	.08
08	.55	.06	.68	.12	.54	.16	1.71	.77	.19	1.07
09	.80	.32	.63	.02	.06	1.07	.43	1.10	.04	.71
10	.55	.55	.09	.04	.40	.33	.04	.89	.14	.03
11	.21	.97	.48	.36	.11	.03	.08	.02	.06	.15
12	1.51	.01	.99	.27	.37	.19	.97	.59	.12	.09
13	.58	.41	.15	.11	.21	.34	.17	.01	.36	.43
14	2.08	.92	.19	.38	1.02	.28	.01	.07	.31	.39
15	.36	.70	.98	.03	.04	.62	.12	.09	.03	.10
16	.20	.34	.18	.19	.19	.68	.50	.70	1.03	.69
17	.07	.33	.76	.83	.84	.96	.34	.22	.02	.30
18	.63	.29	.15	.23	.69	.92	.16	.32	.01	.17
19	1.93	1.53	.95	.44	.51	.00	1.09	.01	.17	.65
20	.11	.07	.76	.37	.31	.24	.31	.97	.27	.02
21	.69	.18	.08	.03	.08	.60	.51	.48	.66	.06
22	.07	.26	1.56	.74	.90	.07	.53	.12	.15	.11
23	.24	.21	.30	.17	.32	.71	.32	.31	.89	1.11
24	.04	.13	.99	.03	.21	.08	.07	.20	.01	.01
25	.55	.70	.12	.03	.27	.76	.01	.20	.08	.61

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Table 29. WK Unstratified Sample DIF_{Lord}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.09	1.13	.22	.41	.20	.56	1.59	1.91	1.11	.10
02	.21	.38	.86	.55	.37	.03	.00	.53	.12	.27
03	.73	1.50	1.47	1.03	.30	.24	.24	.42	.54	.35
04	.40	2.86	.55	.08	1.27	.26	.07	.75	.11	.01
05	.02	1.30	.34	.24	.15	.10	.29	.17	.11	.18
06	.15	.66	1.21	.29	.93	.44	1.53	.01	.40	.37
07	.32	.34	.38	.32	.02	.58	.28	.28	.52	.15
08	.57	.27	.70	.05	.57	.30	1.59	.68	.34	1.09
09	.76	.38	.76	.18	.14	.98	.45	.98	.06	.70
10	.54	.63	.28	.24	.24	.65	.51	.90	.31	.12
11	.23	.97	.58	.42	.05	.35	.58	.11	.44	.03
12	1.63	.42	.97	.15	.46	.05	1.20	.55	.37	.04
13	.58	.19	.13	.15	.14	.30	.06	.22	.38	.48
14	1.86	.97	.06	.14	.90	.37	.01	.12	.25	.44
15	.34	1.08	1.09	.13	.09	.51	.14	.20	.14	.04
16	.19	.20	.11	.18	.17	.80	.38	.81	.96	.76
17	.05	.61	.83	.97	.85	.88	.31	.04	.04	.28
18	.68	.05	.20	.35	.63	.97	.10	.64	.08	.13
19	1.90	1.65	.91	.43	.55	.17	.95	.10	.01	.58
20	.13	.11	.53	.19	.20	.11	.26	.78	.12	.11
21	.81	.19	.21	.15	.14	.64	.56	.68	.64	.08
22	.10	.17	1.41	.66	.85	.10	.65	.08	.13	.07
23	.24	.36	.29	.13	.38	.82	.11	.36	.74	1.13
24	.04	.06	.69	.11	.21	.20	.03	.07	.09	.02
25	.62	.69	.26	.03	.27	1.04	.67	.70	.35	.71

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Aguilar (2017) and Carretta et al. (2016) did not include either CTT or IRT subgroup difference analyses in their report. Morath et al. (2008) included CTT Cohen's *d* results for Form S. The authors did not observe any items with large effect sizes on Form S. There were a few items with moderate effect sizes as well.

3.3.9 Conclusion

Overall, the items in the WK subtest are of below average difficulty and of moderate-to-high discriminability. Item omission rates and longstring option endorsement across the last five items do not suggest speededness or overt guessing. The removal of any one of the items would have a small effect on the internal consistency of the subtest. The internal consistency is fairly high which may suggest that the measure is redundant. No distractors on T1 and one on T2 were chosen more frequently than the correct response. One distractor on T1 and none on T2 were chosen infrequently. Common items analysis showed small differences in item difficulty between Forms T1 and T2. Most items showed small-to-moderate (no large) Cohen's *d* among the subgroups.

Similarly, most items showed small-to-moderate DIF among the subgroups, with a few items showing large DIF for F/M, B/W, and A/W subgroup comparisons.

3.4 Math Knowledge (MK)

MK is the fourth cognitive subtest presented to test-takers and is one of the five subtests that make up the Officership portion of the AFOQT. It consists of 25 items with five multiple choice options. Test-takers are allotted 22 minutes to complete the test, for an average of 52.8 seconds per item.

3.4.1 Descriptives

Table 30 summarizes the means, standard deviations, skewness, and kurtosis for the unstratified sample across Forms T1 and T2. The means ranged from .22 to .77 with an average of .57 on T1 and from .22 to .80 with an average of .57 on T2. The means generally decreased throughout the test. The means for the first five items exceeded .60, whereas the means for the last five items fell below .50 across Forms T1 and T2. Skewness tended to switch from negative to positive throughout the test, with the values remaining within the acceptable range.

Table 30. MK Unstratified Sample Descriptives

Item #	T1				T2			
	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
01	.63	.48	-.54	-1.71	.74	.44	-1.08	-.83
02	.73	.45	-1.02	-.97	.71	.46	-.91	-1.17
03	.68	.47	-.76	-1.42	.60	.49	-.42	-1.82
04	.75	.43	-1.15	-.67	.71	.46	-.90	-1.19
05	.68	.47	-.75	-1.44	.67	.47	-.71	-1.49
06	.76	.43	-1.19	-.59	.80	.40	-1.49	.23
07	.77	.42	-1.29	-.32	.65	.48	-.65	-1.58
08	.65	.48	-.62	-1.61	.64	.48	-.59	-1.65
09	.72	.45	-.96	-1.08	.64	.48	-.56	-1.69
10	.69	.46	-.84	-1.29	.62	.49	-.50	-1.75
11	.55	.50	-.19	-1.97	.65	.48	-.61	-1.63
12	.66	.47	-.69	-1.53	.67	.47	-.71	-1.49
13	.55	.50	-.21	-1.96	.77	.42	-1.31	-.30
14	.64	.48	-.58	-1.66	.53	.50	-.13	-1.98
15	.57	.49	-.30	-1.91	.64	.48	-.57	-1.68
16	.43	.49	.29	-1.92	.64	.48	-.59	-1.65
17	.55	.50	-.21	-1.95	.52	.50	-.08	-1.99
18	.49	.50	.06	-2.00	.58	.49	-.34	-1.89
19	.57	.49	-.30	-1.91	.44	.50	.25	-1.94
20	.40	.49	.40	-1.84	.37	.48	.54	-1.71
21	.45	.50	.19	-1.96	.34	.47	.66	-1.57
22	.32	.47	.75	-1.44	.38	.49	.49	-1.76
23	.32	.47	.75	-1.44	.36	.48	.60	-1.64
24	.22	.42	1.33	-.23	.30	.46	.87	-1.24
25	.39	.49	.45	-1.80	.22	.41	1.36	-.15

Note. *M* = Mean; *SD* = Standard Deviation. $N_{T1} = 36,093$; $N_{T2} = 32,944$.

The results for the stratified samples were similar to the unstratified sample in terms of decreasing means and skewness switching from negative to positive throughout the test. USAFA had the highest average mean than the rest of the accession sources, with T1 averaging at .74 and T2 averaging at .72. Carretta et al. (2016), Aguilar (2017), and Morath et al. (2008) do not report item-level central tendency statistics.

3.4.2 Difficulty

Table 31 summarizes the *p*-values and 2PL *b* parameters for the unstratified sample across Forms T1 and T2. The *p*-values ranged widely on both Forms T1 and T2, but in general, they tended to decrease throughout the test. None of the items on T1 showed a ceiling or floor effect. One item on T2 showed a ceiling effect (T2-06 *p*-value = .80). The 2PL *b* parameters tended to switch from negative to positive throughout the test, while staying within the acceptable range.

Table 31. MK Unstratified Sample Difficulty

Item #	T1		T2	
	<i>p</i> -values	2PL <i>b</i>	<i>p</i> -values	2PL <i>b</i>
01	.63	-.49	.74	-1.05
02	.73	-1.04	.71	-.74
03	.68	-.94	.60	-.50
04	.75	-1.64	.71	-.83
05	.68	-.68	.67	-1.45
06	.76	-1.26	.80	-1.73
07	.77	-1.59	.65	-.80
08	.65	-.83	.64	-.77
09	.72	-1.00	.64	-.69
10	.69	-.82	.62	-.82
11	.55	-.23	.65	-1.16
12	.66	-1.32	.67	-1.31
13	.55	-.28	.77	-1.43
14	.64	-.70	.53	-.18
15	.57	-.46	.64	-.71
16	.43	.57	.64	-1.05
17	.55	-.24	.52	-.11
18	.49	.04	.58	-.36
19	.57	-.33	.44	.18
20	.40	.40	.37	.69
21	.45	.19	.34	.85
22	.32	.81	.38	.55
23	.32	.75	.36	.68
24	.22	1.16	.30	.89
25	.39	.46	.22	1.25

Note. *p*-value = Difficulty; 2PL *b* = IRT Difficulty Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing *p*-values and 2PL *b* parameters switching from negative to positive throughout the test. Form the CTT perspective, USAFA scored higher on average than any other accession sources, with T1 averaging at .74 and T2 averaging at .72. There were 11 items on T1 and eight on T2 for

USAFA that reached the ceiling effect. Ceiling effect was also observed for T2-06 for OTS-CIV and ROTC. There were a few items across the accession sources that reached floor effect (e.g., T1-24 for OTS-AD, AECF, ANG, and AFRES; T2-24 for OTS-AD and ANG; and T2-25 for OTS-AD, AECF, ANG, and AFRES). Notably, most items with ceiling effect were towards the beginning of the test and all items with floor effect were towards the end of the test across the accession sources. From the IRT perspective, nine items on T1 and nine on T2 had 2PL b parameters below 2.00 for USAFA.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and by Aguilar (2017) for Form T. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seems to be more difficult. The mean p -values for Forms T1 and T2 were .57 and .57, respectively. The mean p -values for Form S1 and S2 were .64 and .63, respectively.

3.4.3 Discriminability

Table 32 summarizes the ITC's and 2PL a parameters for the unstratified sample across Forms T1 and T2. ITC's ranged from .31 to .57 with an average of .45 for T1 and from .29 to .59 with an average of .43 on T2. Eight items on T1 and two on T2 had unacceptable 2PL a parameters.

Table 32. MK Unstratified Sample Discriminability

Item #	T1		T2	
	ITC	2PL a	ITC	2PL a
01	.57	1.77	.45	1.31
02	.45	1.22	.56	2.01
03	.42	.96	.47	1.08
04	.34	.75	.50	1.51
05	.54	1.64	.29	.51
06	.42	1.11	.35	.94
07	.35	.88	.43	.96
08	.41	.87	.41	.91
09	.44	1.18	.44	.97
10	.49	1.36	.35	.67
11	.49	1.15	.31	.56
12	.31	.55	.31	.58
13	.44	.94	.39	1.04
14	.44	1.01	.47	1.02
15	.38	.74	.43	.96
16	.32	.54	.33	.61
17	.52	1.29	.49	1.15
18	.50	1.19	.53	1.37
19	.53	1.36	.59	1.78
20	.52	1.24	.41	.90
21	.49	1.11	.41	.88
22	.46	1.12	.46	1.05
23	.49	1.24	.46	1.06
24	.49	1.47	.48	1.19
25	.50	1.19	.47	1.34

Note. ITC = Item-Total Correlation; 2PL a = IRT Discriminability Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample, such that there was a wide range of ITC's among the items. 2PL a parameters were not as consistent across the stratified samples. The majority of the items with acceptable discriminability parameters were observed for OTS-CIV and ROTC for Forms T1 and T2. For the rest of the accession sources, the majority of the items had unacceptable 2PL a parameters.

In general, the current findings are consistent with the results reported by Carretta et al. (2016) and by Aguilar (2017) for Form T. When comparing the present findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seemed to be more discriminable ($T1 = .45$ and $T2 = .43$ vs. $S1 = .40$ and $S2 = .39$).

3.4.4 Internal Consistency if Item is Deleted

Table 33 summarizes Cronbach's alphas if an item is deleted for the unstratified sample across Forms T1 and T2. For reference, the MK subtest-level alphas for T1 and for T2 were .84 and .82, respectively. The item-level alphas were high, ranging from .83 to .84 on T1 and from .81 to .82 on T2. All alphas were within the acceptable range. The removal of any one of the items would have a small effect on the internal consistency of the subtest.

Table 33. MK Unstratified Sample Internal Consistency if Item is Deleted

Item #	T1 α	T2 α
01	.83	.81
02	.83	.81
03	.83	.81
04	.84	.81
05	.83	.82
06	.83	.82
07	.84	.81
08	.83	.81
09	.83	.81
10	.83	.82
11	.83	.82
12	.84	.82
13	.83	.81
14	.83	.81
15	.84	.81
16	.84	.82
17	.83	.81
18	.83	.81
19	.83	.81
20	.83	.81
21	.83	.81
22	.83	.81
23	.83	.81
24	.83	.81
25	.83	.81

Note. α = Cronbach's Alpha.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the alphas were all within the acceptable range and no one item was indicated as dramatically improving internal consistency if deleted. OTS-CIV, ROTC, and AFRES on T1 and OTS-CIV and ROTC on T2 were the accession sources with slightly higher alphas (in the .80 range) compared to the rest of the accession sources (in the .70 range) across Forms T1 and T2.

Although the reports by Aguilar (2017), Carretta et al. (2016), and Morath et al. (2008) did not report the values for alpha if items were deleted, the Cronbach's alpha they reported for the subtests were similar to the alphas observed in this research.

3.4.5 Omission Rates and Overt Guessing

Table 34 summarizes the omission rates in a form of missing values for each item and for a grouping of five items for Forms T1 and T2. The longstring omission rate across the last five items was 1.23% for T1 and 1.83% for T2. In general, the omission rates increased throughout the test. For T1, the average omission rate on the first five items was .59%, whereas the average on the last five items was 3.00%. For T2, these averages went from .73% to 3.84%. The longstring overt guessing for the last five items was 2.16% for T1 and 2.47% for T2.

Table 34. MK Unstratified Sample Omission Rates

Item #	T1		T2	
	For Each Item	Across Five Items	For Each Item	Across Five Items
01	.05%		.50%	
02	1.04%		.07%	
03	.48%	.59%	.87%	.73%
04	.62%		.54%	
05	.75%		1.67%	
06	.35%		.27%	
07	.30%		.69%	
08	.21%	.45%	1.25%	.81%
09	1.04%		.56%	
10	.36%		1.30%	
11	1.00%		.90%	
12	.69%		.19%	
13	.83%	1.02%	.35%	.84%
14	1.21%		1.01%	
15	1.36%		1.74%	
16	.82%		.61%	
17	1.09%		2.20%	
18	1.82%	1.33%	1.02%	1.82%
19	.91%		2.53%	
20	2.02%		2.73%	
21	1.73%		3.30%	
22	3.31%		3.02%	
23	3.23%	3.00%	3.89%	3.84%
24	3.11%		4.50%	
25	3.59%		4.47%	

Note. For Each Item = Percent of Missing Values in Each Item; Across Five Items = Percent of Missing Values Across a Grouping of five Items.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the omission rates tended to increase throughout the test. USAFA cadets had the lowest omission rate in the grouping of the last five items (T1: 1.33%; T2: 1.51%).

In general, the current pattern of findings is consistent with the results reported by Morath et al. (2008) and by Carretta et al. (2016) for Form T and Form S for the unstratified sample. These reports did not include the omission and guessing rates for the stratified samples. Aguilar (2017) did not report omission rates or overt guessing.

3.4.6 Distractor Analysis

When examining the response frequencies, the correct response was the most frequently chosen response for all items on both Forms T1 and T2, except for T1-24 and T2-25. These items happen to be common between the parallel forms. There were six items on T1 for USAFA for which a distractor was chosen less than 1% of the time. On T2, there was one item for AECP and nine items for USAFA for which at least one distractor was chosen infrequently.

The pattern of results for the stratified samples was similar to the unstratified sample, such that most of the accession sources had at least one item for which a distractor was chosen more often than the correct response. USAFA cadets on T1 and on T2 had no distractors chosen more frequently than the correct response. The same was true of OTS-CIV on T2. The accession sources with the highest number of distractors chosen more often than the correct response was OTS-AD, AECP, and AFRES. For several items across the accession sources, a distractor was chosen less than 1% (e.g., T1-01, T1-02, T1-03, T1-07, T1-10, T1-12 for USAFA; T2-13 for AECP; and T2-01, T2-02, T2-04, T2-06, T2-11, T2-12, T2-13, T2-16, and T2-18 for USAFA).

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) for Form T. Aguilar (2017) did not report the distractor analysis. When comparing the present results against the results for Form S reported by Morath et al. (2008) for the unstratified sample, all items on Form S seemed to have had no items for which the distractor was chosen more often.

3.4.7 Common Items

Nine items were common between Forms T1 and T2. They are displayed in Table 35. The table also displays the results of the independent samples *t*-tests comparing the *p*-values for these common items. As can be seen, seven *t*-tests were statistically significant, but the effect sizes (Cohen's *d*) were trivial, ranging from .02 to .19. The rest of the *t*-tests were trivial as well, with Cohen's *d* ranging from 0 to .01. Note that the *t*-tests were run for unstratified sample only. A difference in *p*-values was also examined for the stratified samples only. The differences were small, ranging from 0 to .01 (see Attachment 1).

Table 35. MK Unstratified Sample Common Items Difficulty

T1 Item #	T2 Item #	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	Sig.	ES
T1-01	T2-02	.63	.48	.71	.46	.00	.16
T1-07	T2-06	.77	.42	.80	.40	.00	.06
T1-12	T2-11	.66	.47	.65	.48	.00	.04
T1-14	T2-15	.64	.48	.64	.48	.75	.00
T1-18	T2-17	.48	.50	.52	.50	.00	.07
T1-19	T2-18	.58	.49	.58	.49	.04	.02
T1-21	T2-22	.45	.50	.38	.49	.00	.14
T1-24	T2-25	.22	.42	.22	.41	.19	.01
T1-25	T2-24	.39	.49	.30	.46	.00	.19

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; Sig. = Significance Levels; ES = Effect Size.

The only technical report that included an examination of the common items was by Morath et al. (2008) for Form S. Similar to the current findings, the authors did not detect sizable differences among the common items on Form S.

3.4.8 Subgroup Differences

Table 36 summarizes the CTT effect sizes (Cohen's d) for the mean score difference among the gender, racial, and ethnic subgroups across Forms T1 and T2.

Regarding Female (F) and Male (M) subgroup comparisons, all effect sizes were small. On T1, the effect sizes ranged from .07 to .30, with M scoring higher than F on all 25 items. On T2, the effect sizes ranged from .06 to .33, with M scoring higher than F on all 25 items.

For the Black (B) and White (W) subgroup comparisons, six items on T1 and six items on T2 showed moderate effect sizes, with the rest of the items showing small effect sizes. On T1, the effect sizes ranged from .11 to .49, with W scoring higher than B on 25 items. On T2, the effect sizes ranged from .08 to .54, with W scoring higher than B on 25 items.

Regarding Asian (A) and White (W) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .01 to .35, with W scoring higher than A on three items. On T2, effect sizes ranged from .02 to .33, with W scoring higher than A on two items.

For the White Hispanic (WH) and White Non-Hispanic (WnH) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .04 to .22, with WnH scoring higher than WH on 25 items. On T2, effect sizes ranged from .02 to .22, with WnH scoring higher than WH on 24 items.

Regarding Black Hispanic (BH) and Black Non-Hispanic (BnH) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .00 to .10, with BnH scoring higher than BH on 11 items. On T2, the effect sizes ranged from .00 to .13, with BnH scoring higher than BH on four items.

Table 36. MK Unstratified Sample Mean Score Subgroup Differences - Cohen's *d*

Item #	T1														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.59	.64	.10**	.49	.65	.33**	.79	.65	<u>.30**</u>	.62	.65	.06**	.51	.50	<u>.01</u>
02	.68	.74	.15**	.62	.75	.27**	.82	.74	<u>.18**</u>	.72	.74	.06**	.62	.63	<u>.02</u>
03	.62	.70	.17**	.50	.72	.49**	.74	.71	<u>.06**</u>	.65	.72	.15**	.52	.50	<u>.04</u>
04	.67	.78	.26**	.63	.78	.35**	.78	.78	<u>.01</u>	.73	.78	.13**	.63	.63	<u>.01</u>
05	.64	.69	.10**	.57	.70	.28**	.80	.69	<u>.24**</u>	.66	.69	.08**	.57	.57	<u>.01</u>
06	.70	.78	.19**	.62	.78	.38**	.80	.78	<u>.05*</u>	.72	.79	.15**	.60	.62	.05
07	.75	.78	.08**	.73	.78	.11**	.83	.78	<u>.12**</u>	.75	.78	.07**	.71	.73	.04
08	.57	.67	.21**	.51	.68	.36**	.67	.68	<u>.03</u>	.60	.69	.18**	.51	.52	.03
09	.69	.73	.07**	.61	.74	.29**	.80	.73	<u>.15**</u>	.70	.74	.08**	.63	.61	<u>.04</u>
10	.66	.71	.10**	.60	.71	.25**	.80	.71	<u>.21**</u>	.69	.71	.04*	.58	.60	.04
11	.49	.57	.17**	.42	.57	.30**	.66	.57	<u>.19**</u>	.52	.57	.09**	.45	.42	<u>.06</u>
12	.62	.68	.13**	.56	.69	.27**	.68	.69	<u>.02</u>	.63	.69	.14**	.54	.57	.05
13	.53	.56	.07**	.46	.57	.21**	.66	.57	<u>.19**</u>	.53	.57	.08**	.47	.47	.00
14	.57	.67	.21**	.48	.67	.40**	.70	.67	<u>.06**</u>	.62	.67	.11**	.51	.49	<u>.05</u>
15	.51	.60	.18**	.47	.59	.26**	.67	.59	<u>.15**</u>	.54	.59	.11**	.45	.46	.02
16	.34	.46	.25**	.29	.46	.33**	.44	.45	<u>.02</u>	.41	.46	.11**	.32	.31	<u>.02</u>
17	.52	.57	.10**	.42	.58	.32**	.66	.58	<u>.18**</u>	.52	.58	.12**	.43	.42	<u>.01</u>
18	.43	.51	.15**	.34	.51	.35**	.61	.51	<u>.20**</u>	.45	.51	.14**	.35	.34	<u>.03</u>
19	.47	.61	.30**	.39	.62	.46**	.64	.61	<u>.05*</u>	.53	.62	.19**	.38	.40	.04
20	.35	.42	.15**	.22	.44	.45**	.47	.44	<u>.07**</u>	.34	.45	.21**	.23	.23	.02
21	.36	.49	.25**	.26	.49	.47**	.51	.49	<u>.04†</u>	.41	.50	.18**	.31	.27	<u>.10*</u>
22	.30	.33	.07**	.23	.34	.23**	.46	.33	<u>.26**</u>	.29	.34	.10**	.25	.24	<u>.03</u>
23	.26	.35	.18**	.20	.35	.31**	.44	.34	<u>.20**</u>	.27	.35	.17**	.20	.20	<u>.00</u>
24	.19	.24	.12**	.12	.23	.27**	.38	.23	<u>.35**</u>	.20	.23	.09**	.13	.13	.01
25	.31	.42	.22**	.22	.43	.43**	.47	.42	<u>.10**</u>	.33	.43	.22**	.24	.23	<u>.04</u>
Item #	T2														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.70	.75	.12**	.62	.76	.32**	.82	.75	<u>.17**</u>	.73	.76	.06**	.68	.62	<u>.13**</u>
02	.68	.72	.09**	.56	.73	.38**	.84	.72	<u>.26**</u>	.70	.73	.07**	.60	.56	<u>.08†</u>
03	.51	.64	.27**	.44	.64	.41**	.65	.63	<u>.04†</u>	.54	.64	.22**	.48	.46	<u>.05</u>
04	.63	.73	.21**	.57	.74	.38**	.77	.73	<u>.09**</u>	.68	.74	.12**	.59	.57	<u>.03</u>
05	.65	.68	.06**	.61	.68	.15**	.71	.68	<u>.08**</u>	.64	.68	.08**	.57	.61	.10*
06	.77	.81	.09**	.74	.81	.18**	.86	.81	<u>.14**</u>	.78	.81	.07**	.75	.74	<u>.01</u>
07	.62	.67	.11**	.52	.69	.38**	.66	.69	.07**	.61	.70	.19**	.53	.53	.01
08	.59	.66	.14**	.51	.67	.32**	.72	.66	<u>.13**</u>	.62	.67	.10**	.56	.51	<u>.10*</u>
09	.57	.66	.19**	.46	.67	.45**	.69	.67	<u>.04†</u>	.61	.67	.14**	.49	.47	<u>.05</u>
10	.55	.64	.19**	.48	.65	.36**	.68	.65	<u>.06**</u>	.58	.65	.15**	.53	.48	<u>.09†</u>
11	.60	.66	.13**	.55	.67	.26**	.65	.67	<u>.03</u>	.63	.67	.10**	.58	.55	<u>.07</u>
12	.64	.68	.08**	.63	.67	.08**	.75	.67	<u>.18**</u>	.68	.66	<u>.02</u>	.64	.63	<u>.01</u>
13	.70	.80	.25**	.65	.80	.37**	.82	.80	<u>.07**</u>	.74	.80	.15**	.70	.66	<u>.08†</u>
14	.46	.56	.20**	.40	.56	.33**	.64	.55	<u>.17**</u>	.50	.56	.13**	.44	.40	<u>.07</u>
15	.58	.66	.16**	.48	.67	.40**	.70	.67	<u>.06**</u>	.61	.67	.13**	.50	.49	<u>.03</u>
16	.56	.67	.23**	.52	.67	.30**	.71	.66	<u>.09**</u>	.63	.67	.07**	.55	.53	<u>.05</u>
17	.46	.54	.15**	.35	.55	.39**	.63	.54	<u>.17**</u>	.48	.55	.14**	.36	.37	.02
18	.47	.62	.33**	.37	.63	.54**	.63	.63	<u>.02</u>	.54	.64	.20**	.40	.39	<u>.03</u>
19	.36	.47	.21**	.24	.47	.47**	.57	.46	<u>.21**</u>	.40	.47	.16**	.28	.25	<u>.07</u>
20	.33	.38	.10**	.29	.38	.19**	.49	.38	<u>.23**</u>	.34	.38	.08**	.27	.29	.04
21	.31	.36	.10**	.20	.37	.36**	.40	.37	<u>.07**</u>	.30	.38	.15**	.22	.21	<u>.02</u>
22	.30	.41	.23**	.22	.42	.41**	.42	.41	<u>.02</u>	.34	.42	.17**	.23	.23	<u>.00</u>
23	.29	.38	.17**	.24	.38	.28**	.47	.37	<u>.21**</u>	.33	.38	.10**	.26	.24	<u>.05</u>
24	.24	.32	.18**	.16	.33	.39**	.37	.33	<u>.10**</u>	.25	.34	.18**	.17	.16	<u>.03</u>
25	.19	.23	.10**	.12	.23	.26**	.36	.23	<u>.33**</u>	.20	.23	.08**	.15	.13	<u>.06</u>

†*p* ≤ .100; **p* ≤ .050; ***p* ≤ .010; unmarked are non-significant *p*.

Note. *M_F* = Mean for Female Test-Takers; *M_M* = Mean for Male Test-Takers; *M_B* = Mean for Black/African American Test-Takers; *M_W* = Mean for White Test-Takers; *M_A* = Mean for Asian Test-Takers; *d* = Cohen's *d*. Underlined the effect sizes for the items for which the minority subgroups scored higher than their majority subgroups.

The general pattern of results was consistent among the stratified and unstratified samples, such that the majority of the items had small effect sizes. Some items had moderate effect sizes. Only two items showed large effect size – T2-05 and T2-21 for A/W for AECP. Similar to the unstratified sample, Asians outscored their White counterparts on many items across the stratifications on Forms T1 and T2. For B/W, the number of items with moderate effect sizes was higher than for any other subgroups across the stratifications.

Table 37 displays the IRT DIF_{MH} results and Table 38 displays the DIF_{Lord} results for the same subgroups. T2-02 and T2-07 showed moderate effect sizes for A/W. T2-12 showed moderate effect size for B/W subgroup comparisons.

Table 37. MK Unstratified Sample DIF_{MH}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.59	.15	.86	.36	.04	.41	.47	.52	.40	.37
02	.11	.09	.22	.12	.16	.81	.66	1.29	.64	.16
03	.03	1.14	.55	.41	.27	.38	.17	.53	.39	.03
04	.65	.68	.72	.40	.08	.11	.43	.16	.11	.08
05	.55	.29	.49	.26	.11	.39	.68	.14	.04	.41
06	.19	.65	.57	.45	.20	.38	.81	.46	.26	.06
07	.36	.67	.12	.01	.17	.41	.13	1.10	.37	.41
08	.27	.54	1.11	.51	.29	.24	.27	.13	.13	.43
09	.53	.01	.07	.03	.12	.02	.21	.45	.10	.08
10	.48	.33	.37	.32	.28	.17	.07	.28	.22	.43
11	.09	.00	.08	.14	.39	.09	.32	.75	.02	.27
12	.01	.32	.83	.33	.08	.32	1.19	.46	.63	.05
13	.53	.39	.06	.14	.12	.47	.11	.16	.22	.15
14	.23	.68	.62	.21	.12	.09	.49	.22	.01	.10
15	.10	.01	.07	.15	.00	.15	.00	.35	.05	.07
16	.62	.63	.95	.13	.04	.42	.14	.03	.09	.12
17	.53	.04	.15	.07	.21	.28	.08	.17	.02	.29
18	.17	.24	.02	.16	.06	.66	.32	.77	.34	.05
19	.67	.87	.87	.49	.50	.08	.04	.31	.02	.04
20	.25	1.04	.92	.54	.25	.45	.97	.40	.15	.27
21	.43	1.10	.96	.40	.55	.46	.05	.51	.21	.08
22	.63	.41	.19	.04	.22	.27	.22	.82	.22	.13
23	.08	.15	.24	.35	.01	.03	.64	.27	.13	.06
24	.41	.17	.61	.16	.08	.07	.17	.59	.32	.18
25	.26	.89	.76	.62	.06	.65	.70	.83	.24	.17

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Table 38. MK Unstratified Sample DIF_{Lord}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.35	.08	1.19	.32	.06	.20	.05	.54	.28	.34
02	.01	.26	.50	.26	.19	.36	.18	1.22	.27	.14
03	.06	.78	.41	.20	.27	.53	.30	.44	.44	.00
04	.66	.19	.66	.14	.09	.30	.19	.06	.06	.08
05	.33	.29	.85	.26	.01	.55	1.04	.20	.21	.46
06	.24	.35	.38	.25	.27	.32	.70	.44	.23	.03
07	.40	1.09	.10	.22	.18	.31	.16	1.11	.38	.40
08	.29	.12	.99	.31	.32	.15	.13	.16	.14	.44
09	.45	.20	.23	.19	.06	.12	.50	.43	.10	.07
10	.34	.45	.67	.41	.29	.13	.06	.30	.12	.47
11	.00	.23	.30	.24	.42	.18	.47	.88	.17	.26
12	.14	.33	.93	.04	.05	.46	1.43	.46	.81	.03
13	.52	.73	.29	.29	.17	.58	.25	.16	.25	.13
14	.27	.34	.40	.00	.09	.20	.12	.29	.03	.10
15	.06	.47	.11	.10	.01	.01	.28	.29	.07	.05
16	.51	.04	1.04	.20	.13	.38	.25	.10	.23	.11
17	.36	.13	.25	.03	.13	.12	.24	.25	.08	.24
18	.08	.07	.35	.04	.06	.85	1.01	.60	.45	.01
19	.78	.66	.51	.35	.45	.23	.86	.46	.16	.09
20	.13	.81	.52	.42	.18	.43	.89	.53	.21	.29
21	.49	.86	.66	.26	.57	.40	.26	.39	.14	.11
22	.55	.54	.60	.12	.31	.38	.57	.66	.21	.13
23	.13	.03	.25	.27	.08	.07	.31	.41	.14	.07
24	.18	.03	1.11	.12	.01	.10	.64	.24	.35	.21
25	.34	.68	.29	.48	.13	.36	.10	1.09	.14	.17

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Aguilar (2017) and Carretta et al. (2016) did not include either CTT or IRT subgroup difference analyses in their reports. Morath et al. (2008) included CTT Cohen's *d* results for Form S. The authors did not observe any items with large effect sizes. However, there were items with moderate subgroup differences, with the greatest number of items observed for B/W comparisons. Our pattern of results for Form T is consistent.

3.4.9 Conclusion

Overall, the items in the MK subtest seem to be of below average difficulty and of moderate-to-high discriminability. Item omission rates and longstring option endorsement across the last five items suggest slight speededness and overt guessing. The removal of any one of the items would have a small effect on the internal consistency of the subtest. The internal consistency is fairly high which may suggest that the measure is redundant. There was one distractor chosen more frequently than the correct response, which happens to be common between Forms T1 and T2. No distractors were chosen infrequently. Common items analysis showed small differences in item difficulty

between Forms T1 and T2. Most items showed small-to-moderate (no large) Cohen's *d* and DIF across the subgroups.

3.5 Reading Comprehension (RC)

RC is the fifth cognitive subtest presented to test-takers and is one of the five subtests that make up the Officership portion of the AFOQT. It consists of 25 items with five multiple choice options. Test-takers are given 38 minutes to complete the test, for an average of 91.2 seconds per item. Note that RC was not included in Form S and therefore the comparisons between the current report and the report by Morath et al. (2008) was not possible.

3.5.1 Descriptives

Table 39 summarizes the means, standard deviations, skewness, and kurtosis for the unstratified sample across Forms T1 and T2. The item means ranged from .37 to .93 with an average of .68 on T1 and from .37 to .91 with an average of .69 on T2. The means tended to decrease throughout the test but not to the same extent as the rest of the AFOQT subtests. The means for the first five items exceeded .60, whereas the means for the last five items fell at or below .65 across Forms T1 and T2. Some items towards the middle or end of the test had fairly high means (e.g., T1-19 *M* = .93; T2-24 *M* = .68). Skewness remained primarily negative throughout the test, with a few items showing positive skewness. While most of the items showed acceptable skewness, one item on T1 (T1-19) showed unacceptable skewness (-3.43).

Table 39. RC Unstratified Sample Descriptives

Item #	T1				T2			
	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
01	.79	.41	-1.40	-.04	.91	.28	-2.90	6.39
02	.82	.38	-1.67	.79	.81	.39	-1.62	.62
03	.83	.37	-1.79	1.19	.72	.45	-.97	-1.06
04	.67	.47	-.71	-1.50	.74	.44	-1.10	-.78
05	.63	.48	-.53	-1.72	.80	.40	-1.47	.17
06	.43	.49	.30	-1.91	.72	.45	-.95	-1.09
07	.75	.43	-1.18	-.60	.91	.29	-2.87	6.24
08	.87	.34	-2.16	2.67	.74	.44	-1.07	-.85
09	.67	.47	-.72	-1.48	.76	.43	-1.20	-.56
10	.84	.37	-1.86	1.44	.90	.30	-2.60	4.77
11	.54	.50	-.17	-1.97	.51	.50	-.05	-2.00
12	.68	.47	-.78	-1.40	.82	.39	-1.65	.72
13	.91	.29	-2.84	6.08	.52	.50	-.10	-1.99
14	.60	.49	-.41	-1.84	.53	.50	-.11	-1.99
15	.51	.50	-.05	-2.00	.65	.48	-.62	-1.61
16	.63	.48	-.55	-1.69	.79	.41	-1.44	.06
17	.37	.48	.54	-1.71	.85	.36	-1.95	1.81
18	.90	.29	-2.74	5.52	.73	.44	-1.05	-.89
19	.93	.25	-3.43	9.78	.58	.49	-.34	-1.88
20	.75	.43	-1.13	-.71	.37	.48	.56	-1.69
21	.53	.50	-.12	-1.99	.49	.50	.06	-2.00
22	.64	.48	-.58	-1.67	.79	.41	-1.44	.08
23	.71	.46	-.90	-1.18	.47	.50	.13	-1.98
24	.51	.50	-.03	-2.00	.68	.47	-.79	-1.38
25	.57	.50	-.26	-1.93	.73	.44	-1.04	-.92

Note. *M* = Mean; *SD* = Standard Deviation. $N_{T1} = 36,089$; $N_{T2} = 32,942$.

The pattern of results for the stratified samples was similar to the unstratified sample. While means decreased somewhat throughout the test, this trend was less pronounced than for the rest of the subtests. There were some items whose mean in the middle or towards the end of the test was higher than other items earlier in the test. Some skewness values were below the acceptable levels. USAFA cadets had the highest number of items with means above .80 on Forms T1 and T2, eight and 11, respectively. Most of the accession sources had at least one item with an unacceptable skewness.

Carretta et al. (2016) and Aguilar (2017) did not report item-level central tendency statistics.

3.5.2 Difficulty

Table 40 summarizes the *p*-values and 2PL *b* parameters for the unstratified sample across Forms T1 and T2. The *p*-values ranged widely on both Forms T1 and T2 with no clear pattern of *p*-values decreasing for items positioned later in the subtests. Floor effect was not observed in either form. Potential ceiling effect was observed for seven items in T1 (i.e., had *p*-values over .80), and for six items in T2. Seven items on T1 and seven items on T2 had extreme 2PL *b* values.

Table 40. RC Unstratified Sample Difficulty

Item #	T1		T2	
	<i>p</i> -values	2PL <i>b</i>	<i>p</i> -values	2PL <i>b</i>
01	.79	-2.32	.91	-2.23
02	.82	-1.56	.81	-3.27
03	.83	-1.43	.72	-2.55
04	.67	-.73	.74	-2.40
05	.63	-.83	.80	-2.09
06	.43	.40	.72	-1.62
07	.75	-3.73	.91	-2.09
08	.87	-2.56	.74	-1.08
09	.67	-1.42	.76	-1.38
10	.84	-2.46	.90	-2.59
11	.54	-.33	.51	-.09
12	.68	-1.04	.82	-1.44
13	.91	-3.22	.52	-.16
14	.60	-.87	.53	-.13
15	.51	-.10	.65	-.64
16	.63	-.59	.79	-1.73
17	.37	.95	.85	-1.88
18	.90	-3.10	.73	-1.08
19	.93	-2.74	.58	-.34
20	.75	-1.63	.37	.74
21	.53	-.21	.49	.09
22	.64	-.61	.79	-1.33
23	.71	-1.04	.47	.18
24	.51	-.03	.68	-.79
25	.57	-.49	.73	-1.08

Note. *p*-value = Difficulty; 2PL *b* = IRT Difficulty Parameter.

The results for the stratified samples were similar to the unstratified sample in terms of the *p*-values showing no clear pattern and 2PL *b* parameters showing a handful of items with extreme values. From the CTT perspective, USAFA cadets scored higher on average than any other accession sources, with T1 averaging at .73 and T2 averaging at .75. There were eight items on T1 and 11 on T2 for USAFA that reached the ceiling effect. Ceiling effects were also observed for the rest of the accession sources across at least four items. Notably, these items appeared throughout the test, unlike the rest of the subtests, for which the items with ceiling effect tended to cluster towards the beginning of the test. No items reached floor effect. From the IRT perspective, at least one item across the accession sources had an extreme 2PL *b* value.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and Aguilar (2017) for Form T.

3.5.3 Discriminability

Table 41 summarizes the ITC's and 2PL *a* parameters for the unstratified sample across Forms T1 and T2. ITC's ranged from .22 to .49 with an average of .37 for T1 and from .25 to .52 with an average of .39 on T2. Eight items on T1 and 11 on T2 had acceptable 2PL *a* parameters.

Table 41. RC Unstratified Sample Discriminability

Item #	T1		T2	
	ITC	2PL α	ITC	2PL α
01	.29	.60	.37	1.33
02	.43	1.24	.25	.47
03	.48	1.59	.25	.38
04	.49	1.22	.28	.46
05	.37	.70	.33	.72
06	.42	.88	.33	.61
07	.22	.31	.40	1.47
08	.32	.82	.48	1.21
09	.32	.53	.42	.98
10	.32	.75	.32	.97
11	.34	.56	.34	.59
12	.41	.84	.47	1.39
13	.26	.79	.37	.69
14	.31	.49	.45	.98
15	.33	.53	.50	1.23
16	.49	1.17	.38	.89
17	.34	.61	.40	1.12
18	.28	.80	.47	1.19
19	.31	1.15	.52	1.36
20	.36	.74	.39	.86
21	.35	.59	.37	.68
22	.48	1.21	.47	1.31
23	.44	1.02	.42	.87
24	.48	1.19	.50	1.25
25	.34	.57	.47	1.17

Note. ITC = Item-Total Correlation; 2PL α = IRT Discriminability Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the ITC's varied somewhat across the accession sources, ranging from the low .30s to low .40s, and 2PL α parameters were mostly below the acceptable threshold.

In general, the current findings are consistent with the results reported by Carretta et al. (2016) and by Aguilar (2017) for Form T.

3.5.4 Internal Consistency if Item is Deleted

Table 42 summarizes Cronbach's alphas if an item is deleted for the unstratified sample across Forms T1 and T2. For reference, the overall alpha for each test form was .73 and .78, respectively. The item-level alphas were acceptable, ranging from .71 to .73 for T1 and from .76 to .78 for T2. All alphas were within the acceptable range. The removal of any one of the items would have a small effect on the internal consistency of the RC subtest.

Table 42. RC Unstratified Sample Internal Consistency if Item is Deleted

Item #	T1 α	T2 α
01	.73	.77
02	.72	.78
03	.72	.78
04	.71	.78
05	.72	.77
06	.72	.77
07	.73	.77
08	.73	.76
09	.73	.77
10	.73	.77
11	.73	.77
12	.72	.77
13	.73	.77
14	.73	.77
15	.73	.76
16	.71	.77
17	.73	.77
18	.73	.77
19	.73	.76
20	.72	.77
21	.73	.77
22	.71	.77
23	.72	.77
24	.71	.76
25	.73	.77

Note. α = Cronbach's Alpha.

The pattern of results for the stratified samples was similar to the unstratified sample. All alphas were within the acceptable range on T2. However, on T1, most of the alphas for AECP and USAFA Test-takers were below the .70 threshold. Several items for OTS-AD and ANG test-takers were below the threshold.

Although the reports by Aguilar (2017) and Carretta et al. (2016) did not report the values for alpha if items were deleted, the Cronbach's alpha they reported for the subtests were similar to the alphas observed in this research.

3.5.5 Omission Rates and Overt Guessing

Table 43 summarizes the omission rates in a form of missing values for each item and for a grouping of five items for Forms T1 and T2. The longstring omission rate across the last five items was .13% for T1 and .12% for T2. In general, the omission rates increased throughout the test but minimally. For T1, the average omission rate on the first five items was .10%, whereas the average on the last five items was .46%. For T2, these averages went from .11% to .42%. The longstring overt guessing for the last five items was .87% for T1 and .18% for T2.

Table 43. RC Unstratified Sample Omission Rates

Item #	T1		T2	
	For Each Item	Across Five Items	For Each Item	Across Five Items
01	.10%		.12%	
02	.09%		.03%	
03	.13%	.10%	.16%	.11%
04	.06%		.16%	
05	.14%		.08%	
06	.09%		.06%	
07	.12%		.15%	
08	.07%	.10%	.18%	.11%
09	.11%		.08%	
10	.09%		.06%	
11	.11%		.08%	
12	.05%		.06%	
13	.05%	.08%	.15%	.10%
14	.06%		.10%	
15	.12%		.11%	
16	.11%		.11%	
17	.13%		.21%	
18	.15%	.15%	.33%	.20%
19	.12%		.11%	
20	.24%		.22%	
21	.20%		.28%	
22	.48%		.23%	
23	.51%	.46%	.45%	.42%
24	.65%		.56%	
25	.47%		.56%	

Note. For Each Item = Percent of Missing Values in Each Item; Across Five Items = Percent of Missing Values Across a Grouping of five Items.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the omission rates tended to increase throughout the test, albeit minimally.

In general, the pattern of findings is consistent with the results reported by Carretta et al. (2016) for Form T. Aguilar (2017) did not report omission rates or overt guessing.

3.5.6 Distractor Analysis

When examining the response frequencies, the correct response was the most frequently chosen response for all items on both Forms T1 and T2, except for T2-20. T1 had 11 items in which a distractor was chosen less than 1% of the time. T2 had 10 items that had response options chosen less than 1% of the time. The pattern of results for the stratified samples was similar to the unstratified sample, such that most of the accession sources selected the correct response more frequently than a distractor. The exception to that was T1-06, for which OTS-CIV and AECF

selected a distractor more frequently than a correct response. Another exception was T2-20, for which all accession sources but USAFA selected a distractor more frequently than the correct response. Many items across the stratification samples had at least one distractor chosen infrequently. In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) for Form T and highlight items and their distractors that may need to be revised. Aguilar (2017) did not report the distractor analysis.

3.5.7 Common Items

RC does not have any common items across Forms T1 and T2.

3.5.8 Subgroup Differences

Table 44 summarizes the CTT effect sizes (Cohen's *d*) for the mean score difference among the gender, racial, and ethnic subgroups across T1 and T2.

For the comparisons between the Female (F) and Male (M) subgroups, all effect sizes were small. On T1, the effect sizes ranged from .01 to .39, with M scoring higher than F on 22 out of 25 items. On T2, the effect sizes ranged from .01 to .28, with M scoring higher than F on 24 out of items.

Regarding Black (B) and White (W) subgroup comparisons, five items on T1 and 12 items on T2 showed moderate effect sizes, with the rest of the items showing small effect sizes. On T1, the effect sizes ranged from .08 to .57, with W scoring higher than B on all 25 items. On T2, the effect sizes ranged from .07 to .67, with W scoring higher than B on all 25 items.

For the Asian (A) and White (W) subgroup comparisons, all effect sizes on T1 were small and one effect size on T2 was moderate (T2-19). On T1, the effect sizes ranged from .02 to .29, with W scoring higher than A on all 25 items. On T2, effect sizes ranged from .07 to .42, with W scoring higher than A on all 25 items.

Regarding White Hispanic (WH) and White Non-Hispanic (WnH) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .02 to .23, with WnH scoring higher than WH on all 25 items. On T2, effect sizes ranged from .04 to .27, with WnH scoring higher than WH on all 25 items.

For the comparisons between the Black Hispanic (BH) and Black Non-Hispanic (BnH) subgroups, all effect sizes were small on both forms. On T1, the effect sizes ranged from .01 to .19, with BnH scoring higher than BH on 10 items. On T2, the effect sizes ranged from .01 to .14, with BnH scoring higher than BH on six items.

Table 44. RC Unstratified Sample Mean Score Subgroup Differences - Cohen's *d*

Item #	T1														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.74	.80	.16**	.72	.80	.19**	.79	.80	.03	.79	.80	.04 [†]	.73	.73	.01
02	.76	.84	.23**	.72	.84	.35**	.80	.84	.12**	.81	.85	.10**	.73	.73	.01
03	.78	.85	.18**	.72	.86	.39**	.81	.86	.14**	.82	.86	.12**	.77	.73	.09*
04	.53	.71	.39**	.46	.72	.57**	.61	.72	.22**	.62	.73	.23**	.49	.47	.03
05	.57	.65	.17**	.52	.65	.28**	.63	.65	.04 [†]	.61	.65	.10**	.56	.52	.07
06	.33	.46	.25**	.30	.46	.33**	.36	.46	.21**	.39	.47	.15**	.29	.31	.04
07	.78	.75	.07**	.73	.76	.08**	.74	.76	.06**	.76	.76	.02	.70	.74	.08 [†]
08	.82	.88	.17**	.81	.88	.21**	.86	.88	.06**	.86	.88	.08**	.79	.82	.07
09	.68	.67	.02*	.56	.70	.30**	.63	.70	.15**	.63	.70	.15**	.57	.57	.02
10	.84	.84	.01	.79	.86	.17**	.82	.85	.10**	.82	.86	.10**	.79	.80	.02
11	.52	.55	.05**	.46	.57	.22**	.48	.57	.18**	.54	.57	.06**	.48	.47	.02
12	.60	.71	.23**	.55	.72	.39**	.59	.72	.29**	.66	.72	.14**	.57	.56	.03
13	.89	.91	.08**	.86	.93	.23**	.87	.92	.19**	.88	.93	.17**	.85	.87	.04
14	.56	.61	.11**	.52	.62	.22**	.55	.62	.15**	.58	.63	.09**	.57	.53	.07
15	.48	.52	.08**	.45	.53	.16**	.47	.53	.13**	.49	.54	.10**	.40	.45	.11*
16	.53	.67	.28**	.44	.68	.50**	.56	.68	.26**	.64	.68	.09**	.49	.47	.06
17	.34	.38	.09**	.26	.40	.30**	.32	.40	.16**	.35	.40	.11**	.25	.27	.05
18	.89	.91	.06**	.86	.92	.20**	.91	.91	.02	.90	.92	.04*	.87	.86	.02
19	.93	.93	.02*	.88	.95	.25**	.93	.94	.07**	.93	.95	.09**	.86	.89	.09*
20	.72	.76	.09**	.65	.77	.30**	.73	.77	.11**	.73	.78	.10**	.68	.66	.05
21	.43	.56	.27**	.42	.56	.29**	.44	.56	.25**	.48	.57	.19**	.42	.44	.03
22	.54	.67	.29**	.45	.68	.50**	.62	.68	.13**	.63	.68	.12**	.52	.45	.13**
23	.66	.72	.14**	.55	.75	.44**	.66	.74	.19**	.70	.75	.11**	.58	.55	.05
24	.37	.56	.38**	.30	.56	.52**	.45	.55	.21**	.49	.56	.14**	.35	.32	.07
25	.53	.58	.10**	.42	.60	.36**	.53	.60	.13**	.56	.60	.08**	.52	.43	.19**
Item #	T2														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.88	.92	.17**	.83	.94	.38**	.87	.93	.23**	.90	.94	.15**	.84	.83	.01
02	.81	.82	.01	.75	.84	.22**	.77	.83	.16**	.81	.84	.08**	.73	.75	.04
03	.69	.73	.08**	.70	.73	.07**	.68	.73	.11**	.71	.73	.05*	.70	.69	.02
04	.68	.76	.20**	.67	.76	.21**	.67	.76	.21**	.73	.77	.09**	.70	.68	.04
05	.72	.82	.26**	.66	.83	.43**	.75	.83	.20**	.80	.83	.08**	.71	.66	.10*
06	.68	.73	.10**	.66	.74	.18**	.63	.74	.26**	.68	.75	.16**	.65	.68	.05
07	.88	.92	.13**	.84	.93	.35**	.85	.93	.33**	.91	.93	.11**	.84	.85	.04
08	.68	.76	.19**	.59	.78	.45**	.64	.78	.34**	.72	.78	.15**	.66	.60	.14**
09	.67	.79	.28**	.64	.80	.39**	.65	.80	.35**	.71	.81	.24**	.65	.64	.02
10	.88	.90	.09**	.82	.92	.33**	.86	.92	.22**	.89	.92	.10**	.83	.82	.02
11	.50	.52	.05**	.45	.53	.17**	.48	.53	.10**	.51	.53	.04 [†]	.47	.46	.02
12	.80	.82	.06**	.70	.85	.42**	.76	.85	.25**	.80	.86	.16**	.74	.71	.08 [†]
13	.54	.52	.05**	.44	.55	.23**	.51	.55	.07**	.48	.56	.14**	.41	.45	.07
14	.48	.54	.11**	.34	.57	.47**	.49	.57	.15**	.49	.58	.17**	.40	.35	.12*
15	.62	.66	.08**	.48	.69	.45**	.61	.69	.18**	.61	.70	.20**	.55	.49	.12
16	.78	.80	.03**	.70	.82	.29**	.77	.81	.10**	.77	.82	.13**	.72	.71	.02
17	.80	.87	.19**	.73	.88	.46**	.80	.88	.23**	.85	.88	.12**	.75	.73	.05
18	.69	.75	.14**	.60	.77	.40**	.68	.77	.22**	.71	.77	.14**	.65	.59	.11*
19	.50	.61	.23**	.33	.65	.67**	.45	.65	.42**	.53	.66	.27**	.40	.35	.11*
20	.34	.38	.08**	.23	.40	.35**	.34	.40	.12**	.36	.40	.08**	.27	.25	.06
21	.47	.49	.04**	.38	.52	.28**	.42	.51	.18**	.46	.52	.12**	.35	.39	.08 [†]
22	.75	.81	.15**	.64	.84	.52**	.71	.83	.32**	.77	.84	.18**	.64	.65	.01
23	.41	.49	.16**	.29	.51	.46**	.40	.51	.23**	.43	.52	.18**	.33	.30	.06
24	.66	.69	.07**	.47	.73	.58**	.61	.73	.26**	.65	.74	.19**	.56	.49	.14**
25	.69	.74	.11**	.56	.77	.49**	.66	.77	.26**	.71	.78	.16**	.64	.58	.13**

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

Note. M_F = Mean for Female Test-Takers; M_M = Mean for Male Test-Takers; M_B = Mean for Black/African American Test-Takers; M_W = Mean for White Test-Takers; M_A = Mean for Asian Test-Takers; d = Cohen's d . Underlined the effect sizes for the items for which the minority subgroups scored higher than their majority subgroups.

The general pattern of results was consistent among the stratified and unstratified samples, such that the majority of the items had small effect sizes. Some items had moderate effect sizes. Only three AECP items showed large effect size – T1-11 for A/W, T2-25 for B/W, and T2-12 for WH/WnH. There were some items where minority test-takers scored higher than their majority counterparts (e.g., T1-07 for OTS-CIV, T2-03 for OTS-AD).

Table 45 displays the IRT DIF_{MH} results and Table 46 displays the DIF_{Lord} results for the same subgroups. T1-04, T1-07, and T1-24 showed moderate effect sizes for F/M. T1-04 also showed a moderate effect size for B/W. T2-03, T2-11, and T2-19 showed moderate to large effect sizes for B/W.

Table 45. RC Unstratified Sample DIF_{MH}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.12	.59	.58	.24	.00	.58	.19	.01	.26	.21
02	.43	.03	.32	.07	.05	.47	.34	.01	.02	.24
03	.21	.13	.37	.06	.36	.08	1.18	.32	.22	.07
04	1.21	1.01	.22	.53	.13	.53	.46	.23	.07	.03
05	.01	.30	.60	.03	.48	.92	.41	.06	.17	.37
06	.43	.03	.27	.21	.28	.05	.84	.39	.23	.35
07	1.12	.82	.20	.21	.49	.24	.27	.53	.13	.39
08	.37	.39	.51	.02	.42	.37	.07	.48	.03	.31
09	.90	.08	.12	.33	.04	.97	.02	.64	.52	.29
10	.93	.57	.25	.06	.03	.02	.08	.21	.09	.14
11	.54	.42	.20	.11	.14	.33	1.08	.41	.47	.12
12	.38	.28	.64	.21	.25	.41	.08	.13	.08	.07
13	.20	.08	.54	.71	.39	.85	.69	.62	.04	.53
14	.22	.36	.08	.02	.31	.05	.42	.41	.02	.33
15	.38	.66	.06	.05	.40	.28	.07	.42	.13	.25
16	.59	.71	.36	.09	.05	.47	.50	.64	.01	.01
17	.31	.13	.14	.14	.41	.53	.44	.01	.12	.11
18	.34	.32	.70	.19	.00	.15	.08	.30	.13	.31
19	.72	.01	.51	.05	.57	.58	1.30	.95	.43	.56
20	.30	.09	.24	.12	.11	.17	.02	.36	.35	.07
21	.55	.08	.55	.48	.19	.38	.41	.11	.04	.33
22	.59	.72	.37	.00	.65	.20	.50	.34	.08	.30
23	.14	.50	.06	.01	.42	.21	.56	.19	.20	.01
24	1.15	1.00	.21	.08	.21	.39	.74	.08	.10	.51
25	.28	.35	.07	.06	.75	.05	.39	.00	.03	.04

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Table 46. RC Unstratified Sample DIF_{Lord}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.10	.66	.67	.40	.03	.64	.55	.39	.45	.18
02	.57	.26	.07	.00	.08	.67	.81	.35	.29	.30
03	.35	.53	.04	.13	.17	.25	1.74	.70	.50	.02
04	1.26	1.20	.39	.54	.08	.41	.92	.11	.37	.01
05	.09	.29	.60	.17	.41	.86	.27	.09	.30	.35
06	.54	.13	.35	.13	.30	.17	1.12	.16	.06	.36
07	1.31	1.30	.47	.52	.55	.32	.35	.96	.17	.35
08	.37	.37	.41	.10	.45	.35	.33	.59	.07	.28
09	1.01	.16	.01	.18	.06	.93	.04	.67	.53	.23
10	.96	.66	.19	.01	.04	.02	.12	.23	.08	.08
11	.60	.61	.15	.27	.14	.49	1.22	.74	.65	.13
12	.42	.28	.72	.13	.21	.34	.31	.19	.15	.04
13	.21	.03	.57	.62	.40	.98	.85	.89	.13	.56
14	.29	.60	.02	.17	.36	.09	.51	.45	.05	.36
15	.42	.91	.14	.13	.39	.30	.27	.30	.21	.29
16	.67	.88	.55	.06	.00	.54	.44	.69	.01	.00
17	.36	.03	.11	.01	.38	.54	.60	.18	.06	.10
18	.36	.27	.69	.26	.06	.10	.08	.08	.00	.33
19	.62	.26	.24	.14	.64	.53	1.50	1.01	.53	.58
20	.32	.13	.20	.01	.08	.24	.02	.53	.39	.03
21	.58	.23	.52	.34	.18	.50	.56	.28	.15	.36
22	.69	.85	.10	.05	.62	.21	.73	.52	.23	.25
23	.09	.57	.21	.01	.38	.15	.54	.02	.14	.02
24	1.21	1.12	.34	.08	.17	.36	.93	.12	.21	.51
25	.31	.16	.15	.24	.81	.08	.52	.15	.12	.09

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

3.5.9 Conclusion

Overall, the items in the RC subtest are dramatically below average difficulty and of low discriminability. Item omission rates and longstring option endorsement across the last five items do not suggest speededness or overt guessing. The removal of any one of the items would have a small effect on the internal consistency of the subtest. No distractors on T1 and one distractor on T2 were chosen more frequently than the correct response. Many distractors were chosen infrequently. Common items analysis showed small differences in item difficulty between Forms T1 and T2. Most items showed small-to-moderate (no large) Cohen's *d* across the subgroups. Similarly, most items showed small-to-moderate DIF across the subgroups, with a few items showing moderate-to-large DIF for B/W.

3.6 Physical Science

Physical Science (PS) is the sixth cognitive subtest presented and is one of the five subtests that make up the Classification portion of the AFOQT. It consists of 20 items with five multiple choice options. Test-takers are given 10 minutes to complete the test, for an average of 30 seconds per item. Note that Form S included GS and Form T included PS. The content coverage of the two subtests is not identical and therefore the comparisons between them are not precise. When comparing GS to PS, we look for general trends rather than for absolute values.

3.6.1 Descriptives

Table 47 summarizes the means, standard deviations, skewness, and kurtosis for the unstratified sample across Forms T1 and T2. The item means ranged from .29 to .84 with an average of .52 on T1 and from .29 to .83 with an average of .55 on T2. The means generally decreased throughout the test. The means for the first five items exceeded .50, whereas the means for the last five items fell below .50 across Forms T1 and T2. Skewness tends to switch from negative to positive throughout the test, with the values remaining within the acceptable range.

Table 47. PS Unstratified Sample Descriptives

Item #	T1				T2			
	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
01	.69	.46	-.81	-1.35	.83	.37	-1.79	1.19
02	.83	.38	-1.74	1.03	.70	.46	-.86	-1.26
03	.84	.36	-1.90	1.63	.78	.41	-1.35	-1.17
04	.72	.45	-.97	-1.06	.58	.49	-.31	-1.90
05	.58	.49	-.32	-1.90	.64	.48	-.60	-1.64
06	.75	.43	-1.17	-.63	.65	.48	-.64	-1.59
07	.51	.50	-.03	-2.00	.45	.50	.20	-1.96
08	.44	.50	.24	-1.94	.61	.49	-.45	-1.80
09	.57	.50	-.27	-1.93	.65	.48	-.65	-1.58
10	.47	.50	.13	-1.98	.63	.48	-.54	-1.71
11	.49	.50	.06	-2.00	.63	.48	-.52	-1.73
12	.54	.50	-.17	-1.97	.59	.49	-.36	-1.87
13	.47	.50	.11	-1.99	.55	.50	-.22	-1.95
14	.34	.47	.69	-1.53	.48	.50	.09	-1.99
15	.44	.50	.24	-1.94	.37	.48	.55	-1.70
16	.40	.49	.41	-1.84	.49	.50	.04	-2.00
17	.48	.50	.07	-1.99	.49	.50	.02	-2.00
18	.39	.49	.47	-1.78	.29	.46	.90	-1.19
19	.31	.46	.80	-1.35	.29	.46	.91	-1.18
20	.29	.45	.91	-1.17	.31	.46	.80	-1.36

Note. *M* = Mean; *SD* = Standard Deviation. $N_{T1} = 36,070$; $N_{T2} = 32,931$.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing means and skewness switching from negative to positive throughout the test. USAFA cadets had the highest mean among accession sources, with T1 averaging at .66 and T2 averaging

at .69. Carretta et al. (2016), Aguilar (2017), and Morath et al. (2008) do not report item-level central tendency statistics.

3.6.2 Difficulty

Table 48 summarizes the p -values and 2PL b parameters for the unstratified sample across Forms T1 and T2. The p -values ranged widely on both Forms T1 and T2, but in general, they tended to decrease throughout the test. Two items on T1 and one on T2 showed a ceiling effect. No items showed a floor effect. The 2PL b parameters tended to switch from negative to positive throughout the test, with T1-03 and T2-03 falling below the acceptable levels.

Table 48. PS Unstratified Sample Difficulty

Item #	T1		T2	
	p -values	2PL b	p -values	2PL b
01	.69	-.82	.83	-1.59
02	.83	-1.56	.70	-.83
03	.84	-2.50	.78	-2.03
04	.72	-.96	.58	-.31
05	.58	-.32	.64	-.80
06	.75	-1.15	.65	-1.44
07	.51	-.05	.45	.19
08	.44	.23	.61	-.46
09	.57	-.40	.65	-1.12
10	.47	.29	.63	-.78
11	.49	.09	.63	-.99
12	.54	-.22	.59	-.82
13	.47	.09	.55	-.27
14	.34	.56	.48	.07
15	.44	.27	.37	.52
16	.40	1.37	.49	.10
17	.48	.06	.49	.05
18	.39	1.15	.29	.92
19	.31	2.95	.29	1.17
20	.29	1.24	.31	2.44

Note. p -value = Difficulty; 2PL b = IRT Difficulty Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing p -values and 2PL b parameters switching from negative to positive throughout the test. From the CTT perspective, USAFA cadets scored higher on average than any other accession source, with the T1 mean of .66 and the T2 mean of .69. There were seven items on T1 and seven on T2 for USAFA that reached the ceiling effect (were too easy). There were a few items across the accession sources that reached a floor effect (e.g., T1-14 and T1-20 for AECP; T2-18 for AECP). Notably, most items with a ceiling effect were towards the beginning of the test and all items with a floor effect were towards the end of the test across the accession sources. From the IRT perspective, one item on T1 (T1-03) and one item on T2 (T2-03) had 2PL b parameters below 2.00 across most of the accession sources. Five items on T1 and 6 on T2 had unacceptable 2PL b parameters for USAFA.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and by Aguilar (2017) for Form T. Recall that the GS subtest in Form S was replaced with the PS subtest in Form T. Therefore, the comparisons made between GS and PS are to be interpreted with caution. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seems to be more difficult. The mean p -values for Forms T1 and T2 were .52 and .55, respectively. The mean p -values for Form S1 and S2 were .66 and .65, respectively.

3.6.3 Discriminability

Table 49 summarizes the ITC's and 2PL a parameters for the unstratified sample across Forms T1 and T2. ITC's ranged from .22 to .59 with an average of .42 for T1 and from .25 to .54 with an average of .41 on T2. Ten items on T1 and eight items on T2 had acceptable 2PL a parameter.

Table 49. PS Unstratified Sample Discriminability

Item #	T1		T2	
	ITC	2PL a	ITC	2PL a
01	.48	1.29	.40	1.30
02	.40	1.31	.49	1.37
03	.29	.75	.33	.68
04	.48	1.30	.54	1.47
05	.55	1.56	.42	.86
06	.45	1.27	.30	.46
07	.39	.68	.52	1.30
08	.52	1.25	.51	1.32
09	.41	.78	.34	.62
10	.32	.47	.40	.78
11	.38	.67	.34	.56
12	.46	.99	.31	.46
13	.52	1.26	.46	.99
14	.59	1.82	.52	1.29
15	.47	1.02	.53	1.40
16	.25	.30	.30	.45
17	.50	1.15	.33	.52
18	.30	.42	.48	1.20
19	.22	.27	.41	.86
20	.39	.80	.25	.33

Note. ITC = Item-Total Correlation; 2PL a = IRT Discriminability Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample, such that there was a wide range of ITC's among the items and the majority of the items had low 2PL a parameters. The only exception to that was ROTC on T1, for which the number of acceptable 2PL a parameters was higher than the number of unacceptable parameters.

In general, the current pattern is consistent with the results reported by Carretta et al. (2016) and Aguilar (2017) for Form T. The mean ITCs for GS Form S1 and S2 were .30 and .27, respectively. The PS subtest on Form T seems to be more discriminating than the GS subtest on Form S.

3.6.4 Internal Consistency if Item is Deleted

Table 50 summarizes Cronbach's alphas if item is deleted for the unstratified sample across Forms T1 and T2. For reference, the PS subtest-level alphas for T1 and for T2 were .75 and .74, respectively. The item-level alphas were high, ranging from .73 to .76 on T1 and from .72 to .74 on T2. All alphas were within the acceptable range. The removal of any one of the items would have a small effect on the internal consistency of the subtest.

Table 50. PS Unstratified Sample Internal Consistency if Item is Deleted

Item #	T1 α	T2 α
01	.74	.73
02	.74	.72
03	.75	.73
04	.74	.72
05	.73	.73
06	.74	.74
07	.75	.72
08	.74	.72
09	.74	.73
10	.75	.73
11	.75	.73
12	.74	.74
13	.73	.72
14	.73	.72
15	.74	.72
16	.76	.74
17	.74	.74
18	.75	.72
19	.76	.73
20	.74	.74

Note. α = Cronbach's Alpha.

The alphas for half of the accession sources were acceptable and for the other half slightly below the acceptable levels. Specifically, the alphas for OTS-CIV, ROTC, ANG, and AFRES on T1 were above the .70 threshold. The acceptable alphas on T2 were observed for OTS-CIV, ROTC, and AFRES. The rest of the accession sources exhibited lower alphas (in the .60 range). Interestingly, the removal of 10 items on T1 for ANG and of eight items on T2 for AFRES would result in alphas crossing the .70 threshold.

Although the reports by Aguilar (2017), Carretta et al. (2016), and Morath et al. (2008) did not report the values for alpha if items were deleted, the values they reported for the subtests were similar to the alphas observed in this research.

3.6.5 Omission Rates and Overt Guessing

Table 51 summarizes the omission rates in a form of missing values for each item and for a grouping of five items for Forms T1 and T2. The longstring omission rate across the last five items was .06% for T1 and .05% for T2. In general, the omission rates seem to increase throughout the test, albeit minimally. For T1, the average omission rate on the first five items was .09%, whereas the average on the last five items was .33%. For T2, these averages went from .08% to .30%. The longstring overt guessing for the last five items was .17% for T1 and .10% for T2.

Table 51. PS Unstratified Sample Omission Rates

Item #	T1		T2	
	For Each Item	Across Five Items	For Each Item	Across Five Items
01	.06%		.05%	
02	.07%		.09%	
03	.05%	.09%	.05%	.08%
04	.10%		.12%	
05	.19%		.08%	
06	.08%		.23%	
07	.19%		.14%	
08	.19%	.17%	.11%	.14%
09	.18%		.05%	
10	.20%		.17%	
11	.16%		.14%	
12	.11%		.11%	
13	.11%	.13%	.08%	.12%
14	.07%		.08%	
15	.21%		.21%	
16	.26%		.24%	
17	.26%		.17%	
18	.34%	.33%	.20%	.30%
19	.36%		.40%	
20	.45%		.49%	

Note. For Each Item = Percent of Missing Values in Each Item; Across Five Items = Percent of Missing Values Across a Grouping of five Items.

Results for the stratified samples were similar to the unstratified sample, such that the omission rates tended to increase throughout the test, but all omission rates were fairly small (below 1%). USAFA cadets had the lowest omission rate in the grouping of the last five items (T1: .07%; T2: .10%).

In general, the findings are consistent with the results reported by Morath et al. (2008) for Form S and Carretta et al. (2016) for Form T for the unstratified sample. Aguilar (2017) did not report omission rates or overt guessing. These reports did not include the omission and guessing rates for the stratified samples.

3.6.6 Distractor Analysis

When examining the response frequencies, the correct response was chosen more frequently than a distractor for all items on both Forms T1 and T2.

The USAFA was the only accession source that had all the keyed response options chosen more frequently than the distractors across for both Forms T1 and T2. ROTC and AFRES chose the keyed response options on T2 only. The rest of the accession sources had at least one item for which a distractor was chosen more often than the correct response. The accession sources with the highest number of distractors chosen more often than the correct response was OTS-AD and AECP. For several items across the accession sources, a distractor was chosen less than 1% of the time. A particularly problematic item on T1 was Item T1-03 with the first distractor being chosen either 1% of the time or not at all. Similarly, for T2-06, the fifth response option was chosen either at 1% or not at all.

Carretta et al. (2016) reported that a distractor was chosen more often than the keyed response for T2-18 and one of the distractors was chosen frequently for T2-19 and T2-20. The current study did not observe a similar pattern likely due to utilizing a larger sample. Aguilar (2017) did not report the distractor analysis. When comparing the present results against those for Form S reported by Morath et al. (2008) for the unstratified sample, all items on Form S seemed to have had no items for which the distractor was chosen more often. The authors mentioned that the examination of the stratified samples revealed that the correct response was the most frequently chosen response for all items except for S1-17 and S1-20 on S1 for USAFA. Because Form T PS and Form S GS differ in their item content, it is difficult to make definitive comparisons between the two research efforts.

3.6.7 Common Items

Seven items were common between Forms T1 and T2. They are displayed in Table 52. The table also displays the results of the independent samples *t*-tests comparing the *p*-values for these common items. As can be seen, four *t*-tests were statistically significant, but the effect sizes (Cohen's *d*) were trivial, ranging from .01 to .13. The rest of the *t*-tests were trivial as well, with Cohen's *d* ranging from 0 to .01. Note that the *t*-tests were run for the unstratified sample only. A difference in *p*-values was also examined for the stratified samples only. The differences were small, ranging from |.01| to |.06| (see Attachment 1).

Table 52. PS Unstratified Sample Common Items Difficulty

T1 Item #	T2 Item #	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	Sig.	ES
T1-01	T2-02	.69	.46	.70	.46	.00	.02
T1-02	T2-01	.83	.38	.83	.37	.10	.01
T1-08	T2-07	.44	.50	.45	.50	.01	.02
T1-09	T2-10	.57	.50	.63	.48	.00	.13
T1-12	T2-13	.54	.50	.55	.50	.01	.02
T1-13	T2-14	.47	.50	.48	.50	.13	.01
T1-20	T2-19	.29	.45	.29	.46	.54	.00

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; Sig. = Significance Levels; ES = Effect Size.

The only technical report that included an examination of the common items was by Morath et al. (2008) for Form S. Similar to the current findings, the authors did not detect sizable differences among the common items on Form S.

3.6.8 Subgroup Differences

Table 53 summarizes the CTT effect sizes (Cohen's d) for the mean score difference among the gender, racial, and ethnic subgroups across Forms T1 and T2.

For comparisons between the Female (F) and Male (M) subgroups, most effect sizes were small, while four effect sizes on T1 and three effect sizes on T2 were moderate. On T1, the effect sizes ranged from .01 to .53, with M scoring higher than F on 20 items. On T2, the effect sizes ranged from .01 to .46, with M scoring higher than F on 19 items.

Regarding the Black (B) and White (W) subgroups, four items on T1 and six items on T2 showed moderate effect sizes, with the rest of the items showing small effect sizes. On T1, the effect sizes ranged from .08 to .69, with W scoring higher than B on 20 items. On T2, the effect sizes ranged from .02 to .67, with W scoring higher than B on 20 items.

For the comparisons between the Asian (A) and White (W) subgroups, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .00 to .29, with W scoring higher than A on 14 items. On T2, effect sizes ranged from .01 to .42, with W scoring higher than A on 16 items.

Regarding the White Hispanic (WH) and White Non-Hispanic (WnH) subgroups, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .01 to .23, with WnH scoring higher than WH on 20 items. On T2, effect sizes ranged from .01 to .27, with WnH scoring higher than WH on 19 items.

For comparisons between the Black Hispanic (BH) and Black Non-Hispanic (BnH) subgroups, all effect sizes were small on both forms. On T1, the effect sizes ranged from .01 to .19, with BnH scoring higher than BH on eight items. On T2, the effect sizes ranged from .00 to .21, with BnH scoring higher than BH on four items.

Table 53. PS Unstratified Sample Mean Score Subgroup Differences - Cohen's *d*

Item #	T1														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.61	.71	.23**	.56	.72	.36**	.67	.72	.12**	.66	.72	.14**	.55	.57	.04
02	.77	.85	.21**	.72	.86	.36**	.85	.85	.01	.81	.86	.14**	.72	.73	.01
03	.84	.85	.02†	.81	.85	.13**	.87	.85	.04†	.85	.85	.01	.80	.81	.02
04	.55	.78	.53**	.48	.77	.69**	.67	.77	.23**	.70	.78	.19**	.56	.50	.12**
05	.50	.61	.22**	.46	.60	.29**	.68	.60	.17**	.56	.60	.09**	.46	.47	.03
06	.72	.76	.10**	.57	.80	.53**	.74	.79	.14**	.74	.80	.12**	.60	.59	.01
07	.34	.57	.45**	.35	.54	.40**	.50	.54	.08**	.48	.55	.13**	.38	.36	.05
08	.33	.48	.30**	.29	.48	.38**	.48	.47	.00	.38	.48	.20**	.34	.29	.11*
09	.44	.61	.34**	.40	.60	.40**	.62	.59	.06**	.57	.60	.06**	.45	.42	.07
10	.34	.51	.35**	.28	.51	.45**	.38	.51	.27**	.45	.51	.13**	.32	.30	.05
11	.39	.52	.25**	.43	.50	.14**	.50	.50	.00	.47	.50	.07**	.39	.43	.08†
12	.50	.56	.12**	.40	.58	.36**	.57	.57	.01	.54	.57	.07**	.44	.41	.06
13	.31	.53	.44**	.32	.51	.38**	.48	.51	.06**	.44	.51	.15**	.37	.33	.09†
14	.17	.40	.48**	.15	.38	.50**	.32	.38	.12**	.30	.39	.18**	.21	.16	.14**
15	.38	.46	.16**	.30	.47	.34**	.48	.47	.02	.40	.47	.14**	.37	.31	.12**
16	.39	.41	.04**	.35	.42	.14**	.37	.41	.08**	.40	.42	.04*	.33	.35	.04
17	.39	.51	.24**	.34	.51	.34**	.54	.51	.06**	.45	.51	.14**	.35	.35	.01
18	.33	.40	.15**	.31	.40	.19**	.39	.40	.03	.37	.40	.07**	.30	.31	.02
19	.26	.33	.15**	.24	.33	.20**	.29	.33	.09**	.28	.34	.13**	.24	.25	.03
20	.19	.33	.30**	.18	.32	.29**	.30	.32	.04†	.26	.32	.13**	.21	.19	.05
Item #	T2														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.78	.85	.18**	.72	.86	.39**	.85	.86	.03	.81	.86	.16**	.73	.73	.00
02	.62	.73	.24**	.57	.73	.37**	.68	.73	.11**	.66	.74	.17**	.57	.58	.01
03	.68	.81	.31**	.69	.80	.27**	.78	.80	.05*	.78	.80	.06**	.72	.70	.05
04	.47	.61	.29**	.37	.62	.51**	.64	.61	.06**	.57	.62	.10**	.42	.38	.07
05	.49	.70	.45**	.47	.68	.43**	.64	.68	.07**	.64	.68	.09**	.58	.47	.21**
06	.61	.67	.13**	.53	.68	.31**	.64	.68	.08**	.65	.68	.07**	.58	.54	.07
07	.34	.49	.30**	.27	.49	.46**	.47	.49	.03	.42	.49	.14**	.33	.28	.11*
08	.54	.63	.19**	.41	.65	.50**	.63	.65	.04†	.62	.64	.05**	.53	.43	.20**
09	.67	.65	.03*	.54	.69	.32**	.68	.68	.01	.64	.69	.11**	.58	.53	.10*
10	.50	.68	.37**	.46	.66	.42**	.69	.66	.06*	.62	.66	.08**	.53	.47	.12**
11	.51	.67	.32**	.51	.65	.30**	.63	.65	.05*	.61	.66	.10**	.57	.51	.13**
12	.51	.62	.22**	.49	.61	.25**	.55	.61	.12**	.58	.61	.08**	.52	.49	.05
13	.50	.57	.13**	.42	.58	.34**	.56	.58	.03	.53	.59	.11**	.47	.42	.09†
14	.31	.54	.46**	.32	.51	.38**	.48	.51	.06**	.46	.51	.11**	.36	.32	.09†
15	.27	.40	.26**	.19	.41	.46**	.37	.40	.08**	.34	.41	.15**	.24	.20	.09
16	.44	.51	.14**	.48	.49	.02	.52	.49	.06*	.49	.49	.01	.46	.48	.03
17	.43	.52	.18**	.40	.52	.24**	.48	.52	.07**	.49	.52	.06**	.36	.41	.10*
18	.16	.34	.40**	.18	.32	.30**	.34	.31	.05*	.26	.32	.12**	.17	.20	.06
19	.20	.33	.28**	.18	.32	.30**	.29	.32	.05*	.26	.32	.13**	.20	.19	.04
20	.24	.34	.22**	.23	.33	.22**	.30	.33	.08**	.31	.33	.06**	.28	.24	.10*

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

Note. M_F = Mean for Female Test-Takers; M_M = Mean for Male Test-Takers; M_B = Mean for Black/African American Test-Takers; M_W = Mean for White Test-Takers; M_A = Mean for Asian Test-Takers; d = Cohen's d . Underlined the effect sizes for the items for which the minority subgroups scored higher than their majority subgroups.

The general pattern of results was consistent among the stratified and unstratified samples, such that no large effect sizes were observed, except for two items on Forms T1 and T2 for AECP (T1-10 for A/W and T2-01 for A/W). As discussed earlier, the results for the AECP are to be interpreted with caution due to small sample sizes. The majority of the rest of the effect sizes were small and some were moderate. Generally, the highest number of items with moderate effect sizes was

observed for F/M and for B/W. ROTC had the highest number of items with moderate effect sizes for the B/W (10 on T1 and 10 on T2).

Table 54 displays the IRT DIF_{MH} results and Table 55 displays the DIF_{Lord} results for the same subgroups. There seemed to be a high degree of convergence between the two DIF approaches. T1-03 and T1-07 consistently produced moderate effect sizes for the F/M; T1-04 showed large effect sizes for F/M, B/W, and A/W; T1-05 and T1-10 showed moderate effect sizes for A/W; and T1-14 showed large effect sizes for F/M and moderate effect sizes for B/W. T2-05, T2-09, T2-14, and T2-18 showed moderate effect sizes for F/M; T2-09 showed moderate-to-large effect sizes for F/M and T2-16 showed moderate-to-large effect sizes for B/W.

Table 54. PS Unstratified Sample DIF_{MH}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.10	.02	.56	.03	.41	.21	.28	.21	.43	.53
02	.05	.31	.15	.23	.20	.07	.19	.24	.25	.66
03	1.04	.68	.49	.47	.33	.69	.16	.01	.13	.10
04	1.63	1.68	1.22	.30	.52	.10	.57	.87	.13	.16
05	.29	.74	1.33	.42	.48	1.20	.38	.05	.10	.70
06	.80	.98	.67	.08	.10	.28	.23	.20	.05	.17
07	1.18	.54	.31	.09	.08	.16	.50	.21	.02	.30
08	.22	.07	.14	.30	.48	.49	.49	.19	.38	.29
09	.58	.44	.46	.28	.23	1.34	.20	.22	.22	.09
10	.76	1.12	1.28	.14	.13	.80	.43	.63	.05	.12
11	.13	.85	.09	.21	.38	.61	.05	.01	.07	.24
12	.75	.05	.07	.30	.10	.17	.11	.38	.01	.06
13	1.06	.04	.28	.01	.44	.65	.15	.19	.06	.09
14	1.64	1.04	.76	.10	.49	1.13	.04	.03	.18	.15
15	.50	.03	.24	.06	.29	.09	.71	.09	.14	.15
16	.80	.41	.37	.18	.13	.29	1.15	.51	.43	.23
17	.09	.12	.45	.06	.17	.15	.22	.14	.09	.68
18	.31	.28	.13	.18	.11	1.10	.15	.70	.03	.38
19	.20	.03	.49	.24	.04	.42	.12	.00	.19	.01
20	.54	.07	.24	.11	.08	.33	.15	.23	.04	.13

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Table 55. PS Unstratified Sample DIF_{Lord}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.14	.07	.42	.15	.41	.22	.44	.03	.56	.37
02	.06	.30	.19	.33	.25	.06	.10	.41	.37	.48
03	1.24	.97	.60	.53	.39	.44	.33	.10	.14	.12
04	1.51	1.67	1.08	.44	.42	.12	.84	.58	.05	.00
05	.23	.36	1.28	.16	.41	1.01	.40	.17	.07	.76
06	.79	.96	.56	.16	.14	.68	.23	.24	.19	.16
07	1.03	.26	.16	.05	.02	.19	.70	.03	.12	.29
08	.25	.20	.27	.43	.45	.41	.74	.03	.21	.40
09	.42	.22	.63	.29	.19	1.60	.18	.14	.13	.03
10	.50	.64	1.26	.02	.12	.60	.32	.57	.06	.18
11	.06	1.17	.25	.26	.39	.32	.27	.05	.01	.26
12	.77	.01	.19	.24	.11	.22	.55	.47	.12	.02
13	.84	.02	.05	.08	.43	.71	.09	.03	.00	.02
14	1.59	1.41	.48	.35	.48	1.10	.21	.13	.01	.19
15	.55	.04	.36	.12	.27	.05	.97	.24	.27	.17
16	1.27	1.22	.19	.44	.01	.70	1.83	.54	.60	.32
17	.28	.26	.58	.05	.08	.48	.62	.19	.21	.72
18	.62	.87	.07	.35	.15	1.13	.02	.50	.15	.41
19	.57	.79	.31	.02	.11	.28	.01	.10	.20	.03
20	.43	.07	.00	.11	.06	.17	.60	.24	.26	.07

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Aguilar (2017) and Carretta et al. (2016) did not include either CTT or IRT subgroup difference analyses in their report. Morath et al. (2008) included CTT Cohen's *d* results for Form S. The authors mentioned that an examination of the unstratified sample found one item on S1 with large effect sizes for B/W.

3.6.9 Conclusion

Overall, the items in the PS subtest are of average difficulty and of moderate discriminability. Item omission rates, and longstring option endorsement across the last five items do not suggest speededness or overt guessing. The removal of any one of the items would have a small effect on the internal consistency of the subtest. No distractors were chosen more frequently than the correct response. One distractor on T1 and one on T2 were chosen infrequently. Common items analysis showed small differences in item difficulty between Forms T1 and T2. Most items showed small-to-moderate (no large) Cohen's *d* across the subgroups. Similarly, most items showed small-to-moderate DIF across the subgroups, with a few items showing moderate-to-large DIF for F/M and B/W.

3.7 Table Reading

Table Reading (TR) is the seventh cognitive subtest and is one of the five subtests that make up the Classification portion of the AFOQT. It consists of 40 items with five multiple choice options. Test-takers are allotted 7 minutes to complete the test, for an average of 10.5 seconds per item.

3.7.1 Descriptives

Table 56 summarizes the means, standard deviations, skewness, and kurtosis for the unstratified sample across Forms T1 and T2. The means ranged from .16 to .97 with an average of .68 on T1 and from .15 to .97 with an average of .67 on T2. The means generally decreased throughout the test. The means for the first five items exceeded .80, whereas the means for the last five items fell at or below .40 across Forms T1 and T2. Skewness tended to switch from negative to positive throughout the test. Eleven items on T1 and six on T2 had extreme skewness.

Table 56. TR Unstratified Sample Descriptives

Item #	T1				T2			
	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
01	.95	.22	-4.14	15.13	.90	.31	-2.59	4.69
02	.91	.28	-2.92	6.53	.89	.31	-2.58	4.63
03	.95	.22	-4.18	15.51	.86	.35	-2.05	2.19
04	.93	.25	-3.41	9.65	.92	.26	-3.22	8.39
05	.94	.24	-3.64	11.26	.92	.28	-3.03	7.15
06	.89	.31	-2.49	4.22	.95	.22	-4.18	15.47
07	.93	.26	-3.34	9.18	.97	.18	-5.12	24.23
08	.97	.17	-5.46	27.82	.90	.30	-2.72	5.40
09	.97	.18	-5.08	23.84	.85	.36	-1.92	1.68
10	.95	.23	-3.92	13.36	.95	.21	-4.31	16.57
11	.90	.30	-2.61	4.80	.90	.29	-2.76	5.60
12	.95	.22	-4.07	14.58	.91	.28	-2.97	6.80
13	.90	.29	-2.76	5.59	.92	.26	-3.21	8.28
14	.93	.26	-3.24	8.47	.91	.28	-2.96	6.77
15	.93	.26	-3.27	8.68	.87	.34	-2.14	2.60
16	.87	.34	-2.16	2.66	.90	.30	-2.69	5.21
17	.90	.31	-2.59	4.72	.90	.30	-2.67	5.13
18	.87	.33	-2.22	2.92	.90	.30	-2.62	4.89
19	.88	.33	-2.32	3.38	.84	.37	-1.86	1.45
20	.84	.37	-1.81	1.28	.84	.36	-1.89	1.58
21	.78	.42	-1.32	-.25	.79	.41	-1.39	-.06
22	.83	.38	-1.73	.98	.79	.40	-1.45	.11
23	.73	.44	-1.03	-.95	.70	.46	-.89	-1.20
24	.73	.45	-1.02	-.96	.70	.46	-.85	-1.28
25	.67	.47	-.70	-1.51	.70	.46	-.85	-1.27
26	.70	.46	-.87	-1.24	.67	.47	-.70	-1.51
27	.56	.50	-.23	-1.95	.53	.50	-.10	-1.99
28	.49	.50	.05	-2.00	.48	.50	.08	-1.99
29	.48	.50	.06	-2.00	.43	.50	.28	-1.92
30	.54	.50	-.15	-1.98	.44	.50	.23	-1.95
31	.34	.47	.68	-1.54	.35	.48	.63	-1.60
32	.35	.48	.65	-1.58	.27	.45	1.02	-.97
33	.29	.45	.93	-1.14	.23	.42	1.27	-.40
34	.33	.47	.74	-1.45	.39	.49	.43	-1.81
35	.18	.39	1.66	.74	.18	.39	1.64	.68
36	.17	.38	1.75	1.07	.19	.39	1.60	.56
37	.18	.38	1.66	.76	.15	.35	2.00	2.01
38	.20	.40	1.49	.22	.34	.48	.66	-1.57
39	.16	.36	1.88	1.54	.21	.41	1.44	.08
40	.36	.48	.58	-1.66	.28	.45	.99	-1.02

Note. *M* = Mean; *SD* = Standard Deviation. $N_{T1} = 36,070$; $N_{T2} = 32,929$.

The results for the stratified samples was similar to the unstratified sample in terms of decreasing means and skewness switching from negative to positive throughout the test. The means for all accession sources for Forms T1 and T2 were in the upper .60s lower .70s range.

Carretta et al. (2016), Aguilar (2017), and Morath et al. (2008) do not report item-level central tendency statistics.

3.7.2 Difficulty

Table 57 summarizes the p -values and 2PL b parameters for the unstratified sample across Forms T1 and T2. The p -values ranged widely on both Forms T1 and T2, but in general, they tended to decrease throughout the test. Twenty one items on T1 and 20 on T2 reached a ceiling effect. Five items on T1 and four on T2 reached a floor effect. Sixteen 2PL b parameters on T1 and 15 on T2 were extreme. Generally, the 2PL b parameter tended to decrease throughout the test.

Table 57. TR Unstratified Sample Difficulty

Item #	T1		T2	
	p -values	2PL b	p -values	2PL b
01	.95	-6.18	.90	-4.46
02	.91	-6.09	.89	-4.16
03	.95	-5.24	.86	-4.65
04	.93	-4.65	.92	-4.13
05	.94	-4.65	.92	-3.75
06	.89	-5.57	.95	-3.49
07	.93	-4.12	.97	-3.32
08	.97	-4.38	.90	-3.57
09	.97	-4.13	.85	-4.56
10	.95	-4.26	.95	-3.54
11	.90	-4.13	.90	-3.49
12	.95	-3.40	.91	-3.28
13	.90	-3.19	.92	-2.71
14	.93	-2.90	.91	-2.29
15	.93	-2.49	.87	-2.17
16	.87	-2.12	.90	-1.91
17	.90	-2.00	.90	-1.68
18	.87	-1.59	.90	-1.60
19	.88	-1.68	.84	-1.36
20	.84	-1.27	.84	-1.35
21	.78	-1.02	.79	-.91
22	.83	-1.18	.79	-1.09
23	.73	-.72	.70	-.64
24	.73	-.75	.70	-.60
25	.67	-.54	.70	-.60
26	.70	-.68	.67	-.56
27	.56	-.23	.53	-.11
28	.49	-.05	.48	.03
29	.48	.00	.43	.16
30	.54	-.18	.44	.18
31	.34	.45	.35	.47
32	.35	.45	.27	.78
33	.29	.66	.23	.91
34	.33	.71	.39	.56
35	.18	1.26	.18	1.27
36	.17	1.36	.19	1.46
37	.18	1.58	.15	1.76
38	.20	1.77	.34	1.30
39	.16	2.20	.21	2.24
40	.36	1.44	.28	2.27

Note. p -value = Difficulty; 2PL b = IRT Difficulty Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing but rather high p -values and 2PL b parameters switching from negative to positive throughout the test, with a considerable amount of items having extreme 2PL b parameters. In a similar vein, several items reached a floor effect. Notably, most items with a ceiling effect were towards the beginning and middle of the test and most of the items with a floor effect were towards the end of the test. 2PL b parameters in the first third of the test were extremely low, in the middle third were relatively low, and in the last third were acceptable.

In general, the current findings are consistent with those reported by Carretta et al. (2016) and Aguilar (2017) for Form T. When comparing the present findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seems to be of similar difficulty. The mean p -values for Forms T1 and T2 were .68 and .67, respectively. The mean p -values for S1 and S2 were .67 and .67, respectively.

3.7.3 Discriminability

Table 58 summarizes the ITC's and 2PL a parameters for the unstratified sample across Forms T1 and T2. ITC's ranged from .15 to .69 with an average of .40 for T1 and from .21 to .68 with an average of .41 on T2. Seventeen items on T1 and 16 items on T2 had unacceptable 2PL a parameter.

Table 58. TR Unstratified Sample Discriminability

Item #	T1		T2	
	ITC	2PL α	ITC	2PL α
01	.15	.50	.21	.50
02	.17	.40	.23	.54
03	.18	.60	.22	.40
04	.21	.59	.23	.65
05	.21	.62	.25	.69
06	.18	.39	.26	.95
07	.23	.67	.25	1.17
08	.20	.87	.26	.67
09	.21	.89	.22	.38
10	.22	.72	.25	.96
11	.23	.55	.26	.69
12	.26	.98	.27	.79
13	.28	.78	.31	1.08
14	.29	1.00	.35	1.27
15	.33	1.25	.35	1.00
16	.36	1.05	.40	1.53
17	.39	1.39	.43	1.92
18	.45	1.72	.45	2.08
19	.43	1.65	.47	1.75
20	.51	2.06	.48	1.84
21	.53	1.96	.58	2.70
22	.52	2.28	.51	1.86
23	.62	3.13	.60	2.60
24	.59	2.56	.61	2.85
25	.63	2.89	.60	2.77
26	.59	2.38	.57	2.05
27	.65	2.90	.68	3.60
28	.69	3.43	.62	2.42
29	.58	1.80	.64	2.86
30	.55	1.59	.55	1.73
31	.61	2.27	.59	2.08
32	.60	2.10	.56	1.98
33	.59	2.12	.57	2.23
34	.48	1.26	.42	.92
35	.51	1.78	.50	1.77
36	.49	1.69	.44	1.33
37	.42	1.21	.42	1.32
38	.36	.90	.31	.54
39	.34	.87	.30	.66
40	.25	.41	.25	.44

Note. ITC = Item-Total Correlation; 2PL α = IRT Discriminability Parameter.

Results for the stratified samples were similar to the unstratified sample, such that there was a wide range of ITC's among the items and the majority of the items had strong 2PL α parameters. Notably, low 2PL α parameters were observed mostly in the first third of the test and in a few items at the end of the test across the accessions sources.

In general, the current findings are consistent with the results reported by Carretta et al. (2016) and Aguilar (2017) for Form T. When comparing the present findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seemed to be slightly more discriminable ($T1 = .40$ and $T2 = .41$ vs. $S1 = .36$ and $S2 = .37$).

3.7.4 Internal Consistency if Item is Deleted

Table 59 summarizes Cronbach's alphas if an item is deleted for the unstratified sample across Forms T1 and T2. For reference, the TR subtest-level alphas for T1 and for T2 were .88 and .88, respectively. The item-level alphas were high, ranging from .88 to .89 on T1 and from .88 to .89 on T2. The removal of any one of the items would have a small effect on the internal consistency of the subtest.

Table 59. TR Unstratified Sample Internal Consistency if Item is Deleted

Item #	T1 α	T2 α	Item #	T1 α	T2 α
1	.88	.88	21	.88	.88
2	.88	.88	22	.88	.88
3	.88	.89	23	.88	.88
4	.88	.88	24	.88	.88
5	.88	.88	25	.88	.88
6	.88	.88	26	.88	.88
7	.88	.88	27	.88	.88
8	.88	.88	28	.88	.88
9	.88	.89	29	.88	.88
10	.88	.88	30	.88	.88
11	.88	.88	31	.88	.88
12	.88	.88	32	.88	.88
13	.88	.88	33	.88	.88
14	.88	.88	34	.88	.88
15	.88	.88	35	.88	.88
16	.88	.88	36	.88	.88
17	.88	.88	37	.88	.88
18	.88	.88	38	.88	.89
19	.88	.88	39	.88	.88
20	.88	.88	40	.89	.89

Note. α = Cronbach's Alpha.

The pattern of results for the stratified samples was similar to the unstratified sample, such that the alphas were all high (in the .80s and .90s) and no one item was indicated as dramatically improving internal consistency if deleted.

Although the reports by Aguilar (2017), Carretta et al. (2016), and Morath et al. (2008) did not report the values for alpha if items were deleted, the Cronbach's alpha they reported for the subtests were similar to the alphas observed in this research.

3.7.5 Omission Rates and Overt Guessing

Table 60 summarizes the omission rates in a form of missing values for each item and for a grouping of five items for Forms T1 and T2. The longstring omission rate across for the last five items was 29.91% for T1 and 30.04% for T2. In general, the omission rates increased throughout the test. For T1, the average omission rate on the first five items was .12%, whereas the average on the last five items was 33.42%. For T2, these averages went from .12% to 33.40%. The longstring overt guessing for the last five items was 2.47% for T1 and 2.82% for T2.

Table 60. TR Unstratified Sample Omission Rates

Item #	T1		T2	
	For Each Item	Across Five Items	For Each Item	Across Five Items
1	.06%	.12%	.04%	.12%
2	.08%		.05%	
3	.18%		.11%	
4	.12%		.27%	
5	.18%		.14%	
6	.21%	.17%	.12%	.17%
7	.17%		.14%	
8	.15%		.15%	
9	.15%		.28%	
10	.16%		.16%	
11	.26%	.35%	.19%	.43%
12	.24%		.35%	
13	.31%		.44%	
14	.43%		.50%	
15	.51%		.69%	
16	.82%	1.58%	1.08%	1.82%
17	1.10%		1.11%	
18	1.43%		1.58%	
19	1.93%		2.24%	
20	2.62%		3.11%	
21	3.45%	5.91%	3.81%	6.36%
22	4.35%		4.80%	
23	5.84%		6.13%	
24	7.12%		7.66%	
25	8.81%		9.42%	
26	10.44%	14.82%	11.33%	15.47%
27	12.65%		13.29%	
28	14.74%		15.48%	
29	17.07%		17.49%	
30	19.23%		19.75%	
31	21.75%	25.35%	21.89%	25.94%
32	23.56%		24.13%	
33	25.31%		26.28%	
34	27.22%		28.00%	
35	28.92%		29.39%	
36	30.68%	33.42%	30.91%	33.40%
37	32.24%		32.27%	

38	33.56%	33.54%
39	34.81%	34.66%
40	35.84%	35.61%

Note. For Each Item = Percent of Missing Values in Each Item; Across Five Items = Percent of Missing Values Across a Grouping of five Items.

Results for the stratified samples were similar to the unstratified sample, such that the omission rates tended to increase throughout the test. The accession sources varied in their omission rates across the last five items on both Forms T1 and T2. The lowest and the highest omission rates on T1 were observed for ANG (27.07%) and USAFA (38.60%), respectively. On T2, the lowest omission rate was observed for OTS-AD (30.92%).

In general, the current pattern is consistent with the results reported by Morath et al. (2008) for Form S and by Carretta et al. (2016) for Form T for the unstratified sample. These reports did not include the omission and guessing rates for the stratified samples. Aguilar (2017) did not report omission rates or overt guessing.

3.7.6 Distractor Analysis

When examining the response frequencies, the correct response was chosen more frequently than a distractor for most of the items on both Forms T1 and T2, except for five items on Forms T1 and T2. There were several items, especially towards the beginning of the test, for which at least one distractor was chosen less than 1% of the time.

The stratified samples showed a similar pattern but there were also some differences on T1 towards the end of the subtest. For OTS-CIV and OTS-AD, T1 item 35 through 39 had distractors that were endorsed more frequently than the correct answer. For USAFA and ROTC, item 33 and 35-39 had this issue as well. For AFRES this was an issue for items 37 and 39. For ANG, this occurred for items 36-39. The results for T2 did not show as many of these difference between the unstratified and stratifications towards the end of the subtest.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) for Form T. Aguilar (2017) did not report the distractor analysis. When comparing the present results against the results for Form S reported by Morath et al. (2008) for the unstratified sample, three of the last five items had a distractor option selected more frequently than the correct answer. On S2, two of the last five items had a distractor option selected more frequently than the correct answer.

3.7.7 Common Items

Ten items were common between Forms T1 and T2. They are displayed in Table 61. The table also displays the results of the independent samples *t*-tests comparing the *p*-values for these common items. As can be seen, six *t*-tests were statistically significant, but the effect sizes (Cohen's *d*) were trivial, ranging from .02 to .14. The rest of the *t*-tests were trivial as well, with Cohen's *d* ranging from 0 to .01. Note that the *t*-tests were run for unstratified sample only. A

difference in *p*-values was also examined for the stratified samples only. The differences between the *p*-values were small, ranging from $|.00|$ to $|.07|$ (see Attachment 1).

Table 61. TR Unstratified Sample Common Items Difficulty

T1 Item #	T2 Item #	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	Sig.	ES
T1-05	T2-05	.94	.24	.92	.28	.00	.08
T1-08	T2-07	.97	.17	.97	.18	.01	.02
T1-09	T2-10	.96	.18	.95	.21	.00	.06
T1-14	T2-13	.92	.26	.92	.27	.64	.00
T1-20	T2-19	.84	.37	.84	.37	.28	.01
T1-22	T2-22	.83	.38	.79	.40	.00	.08
T1-25	T2-24	.66	.47	.70	.46	.00	.06
T1-28	T2-27	.49	.50	.53	.50	.00	.07
T1-31	T2-32	.34	.47	.27	.45	.00	.14
T1-35	T2-35	.18	.38	.18	.39	.43	.01

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; Sig. = Significance Levels; ES = Effect Size.

The only technical report that included an examination of the common items was by Morath et al. (2008) for Form S. Similar to the current findings, the authors did not detect sizable differences among the common items on Form S.

3.7.8 Subgroup Differences

Table 62 summarizes the CTT effect sizes (Cohen's *d*) for the mean score difference among the gender, racial, and ethnic subgroups across Forms T1 and T2.

Regarding comparisons between the Female (F) and Male (M) subgroups, all items showed small effect sizes. On T1, the effect sizes ranged from .01 to .11, with M scoring higher than F on all 40 items. On T2, the effect sizes ranged from .01 to .11, with M scoring higher than F on 39 of 40 items.

For the Black (B) and White (W) subgroups, effect sizes were small to moderate on both test forms. On T1, the effect sizes ranged from .17 to .61, with W scoring higher than B on all 40 items. On T2, the effect sizes ranged from .11 to .63, with W scoring higher than B on all 40 items.

Regarding Asian (A) and White (W) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .03 to .11, with W scoring higher than A on 36 items. On T2, effect sizes ranged from .02 to .09, with W scoring higher than A on 30 items.

For the comparisons between the White Hispanic (WH) and White Non-Hispanic (WnH) subgroups, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .07 to .24, with WnH scoring higher than WH on 40 items. On T2, effect sizes ranged from .05 to .21, with WnH scoring higher than WH on 38 items.

Regarding the Black Hispanic (BH) and Black Non-Hispanic (BnH) comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .00 to .09, with BnH scoring higher than BH on 15 items. On T2, the effect sizes ranged from .00 to .08, with BnH scoring higher than BH on 16 items.

Table 62. TR Unstratified Sample Mean Score Subgroup Differences - Cohen's *d*

Item #	T1														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.94	.95	.06**	.93	.95	.11**	.96	.95	.03	.95	.95	.03 [†]	.94	.93	.04
02	.90	.92	.05**	.88	.92	.11**	.92	.91	.02	.90	.92	.06**	.90	.89	.02
03	.94	.95	.04**	.93	.96	.10**	.95	.95	.01	.95	.96	.05**	.93	.93	.01
04	.93	.93	.02 [†]	.92	.94	.06**	.93	.93	.01	.93	.94	.04*	.89	.92	.09*
05	.94	.94	.01	.93	.94	.05**	.94	.94	.00	.93	.94	.04 [†]	.92	.93	.03
06	.88	.89	.04**	.86	.89	.11**	.91	.89	.04 [†]	.87	.90	.08**	.85	.87	.06
07	.92	.93	.05**	.90	.94	.16**	.93	.94	.04	.93	.94	.04*	.89	.90	.02
08	.97	.97	.03**	.95	.97	.15**	.97	.97	.04 [†]	.96	.97	.08**	.96	.95	.01
09	.96	.97	.07**	.94	.97	.15**	.97	.97	.01	.96	.97	.06**	.94	.94	.03
10	.94	.95	.04**	.93	.95	.10**	.95	.95	.01	.94	.95	.06**	.92	.93	.01
11	.88	.90	.07**	.86	.91	.15**	.90	.90	.03	.89	.91	.07**	.85	.86	.03
12	.94	.95	.05**	.91	.96	.21**	.95	.96	.05*	.94	.96	.08**	.93	.91	.05
13	.88	.91	.11**	.85	.92	.23**	.90	.92	.05*	.89	.92	.11**	.86	.85	.02
14	.92	.93	.03**	.88	.94	.20**	.92	.93	.05*	.91	.94	.10**	.90	.89	.04
15	.92	.93	.05**	.86	.94	.31**	.93	.94	.02	.92	.94	.10**	.89	.86	.07
16	.85	.87	.07**	.78	.89	.33**	.87	.89	.04 [†]	.84	.89	.15**	.80	.78	.05
17	.88	.90	.05**	.80	.92	.39**	.90	.91	.06**	.88	.92	.12**	.82	.81	.02
18	.86	.88	.04**	.75	.90	.47**	.88	.90	.07**	.85	.90	.16**	.75	.76	.01
19	.86	.88	.07**	.77	.90	.43**	.89	.90	.05*	.86	.91	.14**	.78	.77	.03
20	.82	.84	.06**	.68	.87	.53**	.85	.87	.06**	.82	.87	.17**	.67	.69	.04
21	.76	.78	.05**	.61	.81	.51**	.79	.81	.05*	.75	.82	.18**	.62	.62	.01
22	.80	.84	.10**	.66	.86	.55**	.84	.86	.07**	.79	.87	.21**	.68	.68	.01
23	.70	.74	.07**	.51	.78	.61**	.74	.77	.08**	.69	.78	.23**	.54	.53	.01
24	.70	.74	.07**	.53	.77	.55**	.74	.77	.07**	.69	.78	.21**	.57	.54	.05
25	.65	.67	.05**	.45	.71	.57**	.68	.71	.07**	.61	.72	.24**	.46	.47	.02
26	.67	.71	.09**	.50	.75	.57**	.71	.74	.08**	.66	.75	.21**	.54	.51	.07
27	.53	.57	.08**	.33	.61	.57**	.55	.61	.11**	.51	.62	.21**	.38	.35	.05
28	.46	.50	.07**	.25	.54	.59**	.48	.54	.11**	.43	.55	.24**	.29	.27	.04
29	.45	.50	.11**	.30	.52	.44**	.48	.52	.09**	.44	.53	.18**	.35	.32	.07
30	.50	.55	.09**	.35	.58	.48**	.52	.58	.11**	.49	.59	.20**	.37	.36	.03
31	.32	.35	.07**	.18	.37	.40**	.34	.37	.06**	.28	.38	.21**	.21	.20	.02
32	.32	.35	.06**	.20	.38	.38**	.34	.38	.07**	.30	.38	.17**	.21	.21	.01
33	.27	.30	.06**	.16	.32	.34**	.29	.32	.07**	.24	.33	.17**	.20	.17	.09*
34	.31	.33	.06**	.22	.35	.28**	.33	.35	.04 [†]	.30	.35	.12**	.24	.22	.04
35	.17	.18	.04**	.09	.20	.29**	.18	.20	.04 [†]	.15	.21	.13**	.10	.10	.00
36	.16	.17	.03*	.09	.19	.24**	.15	.19	.08**	.14	.19	.13**	.10	.10	.01
37	.18	.18	.01	.12	.19	.18**	.17	.19	.05*	.16	.20	.10**	.14	.13	.04
38	.19	.20	.03*	.14	.21	.17**	.20	.21	.04	.19	.22	.07**	.17	.15	.06
39	.15	.16	.01	.11	.17	.17**	.16	.17	.03	.15	.17	.07**	.11	.11	.02
40	.34	.37	.06**	.28	.38	.22**	.36	.38	.04 [†]	.35	.38	.07**	.28	.28	.00

Item #	T2														
	M_F	M_M	d	M_B	M_W	d	M_A	M_W	d	M_{WH}	M_{WnH}	d	M_{BH}	M_{BnH}	d
01	.89	.90	.02	.86	.90	.13**	.91	.90	.04 [†]	.90	.90	.02	.83	.83	.03
02	.90	.89	.00	.88	.90	.06**	.90	.90	.00	.90	.90	.02	.90	.91	.00
03	.84	.86	.07**	.83	.86	.11**	.87	.86	.03	.85	.86	.03	.92	.90	.01
04	.92	.93	.02	.91	.93	.09**	.93	.93	.02	.92	.93	.04*	.93	.93	.02
05	.92	.92	.01	.90	.92	.07**	.93	.92	.02	.91	.92	.05*	.94	.95	.05
06	.95	.95	.02	.93	.95	.12**	.96	.95	.02	.95	.95	.03	.85	.88	.02
07	.97	.97	.00	.95	.97	.09**	.97	.97	.01	.96	.97	.03	.80	.81	.05
08	.90	.91	.03**	.88	.91	.09**	.92	.91	.03	.89	.91	.08**	.91	.93	.08
09	.82	.86	.10**	.81	.85	.13**	.86	.85	.01	.83	.86	.06**	.87	.87	.03
10	.95	.96	.04**	.93	.96	.15**	.95	.96	.02	.96	.96	.01	.86	.88	.11*
11	.89	.91	.06**	.86	.91	.18**	.91	.91	.01	.90	.91	.04*	.87	.89	.01
12	.90	.92	.05**	.87	.92	.19**	.92	.92	.03	.91	.92	.04*	.86	.86	.04
13	.92	.93	.02 [†]	.89	.93	.16**	.93	.93	.01	.92	.93	.03	.80	.79	.06
14	.91	.92	.04**	.85	.93	.29**	.92	.93	.01	.91	.93	.06**	.83	.82	.02
15	.85	.87	.08**	.78	.88	.32**	.87	.88	.04 [†]	.86	.89	.08**	.83	.80	.04
16	.89	.91	.07**	.81	.92	.40**	.91	.92	.04 [†]	.89	.92	.13**	.82	.80	.03
17	.89	.90	.05**	.79	.92	.45**	.91	.92	.05*	.89	.92	.12**	.74	.72	.06
18	.89	.90	.05**	.79	.92	.43**	.91	.92	.02	.89	.92	.11**	.73	.71	.03
19	.82	.85	.07**	.71	.87	.46**	.85	.87	.05*	.82	.87	.15**	.64	.62	.05
20	.83	.85	.07**	.70	.87	.51**	.85	.87	.05*	.83	.88	.13**	.66	.64	.04
21	.76	.80	.10**	.60	.82	.56**	.80	.82	.07**	.76	.83	.18**	.53	.50	.06
22	.77	.80	.08**	.63	.83	.52**	.81	.83	.04 [†]	.77	.83	.18**	.52	.50	.03
23	.68	.71	.09**	.47	.75	.63**	.72	.75	.08**	.68	.75	.16**	.52	.50	.06
24	.69	.70	.03*	.47	.74	.61**	.70	.74	.09**	.66	.75	.21**	.48	.46	.03
25	.68	.70	.06**	.47	.74	.60**	.70	.74	.07**	.66	.75	.19**	.32	.31	.03
26	.63	.68	.11**	.44	.71	.60**	.68	.71	.07**	.63	.72	.20**	.29	.29	.03
27	.50	.54	.07**	.28	.58	.60**	.53	.57	.09**	.48	.58	.21**	.27	.25	.03
28	.45	.49	.07**	.26	.52	.52**	.48	.52	.08**	.44	.53	.17**	.29	.29	.00
29	.40	.44	.08**	.22	.48	.51**	.43	.47	.08**	.39	.48	.18**	.20	.21	.06
30	.42	.45	.06**	.28	.48	.40**	.45	.48	.05*	.41	.48	.16**	.16	.15	.00
31	.33	.36	.07**	.20	.38	.39**	.35	.38	.06**	.31	.39	.16**	.12	.12	.03
32	.26	.28	.05**	.14	.30	.36**	.29	.30	.03	.23	.31	.16**	.27	.28	.04
33	.22	.24	.05**	.12	.26	.34**	.24	.26	.03	.19	.26	.16**	.10	.11	.01
34	.36	.40	.08**	.26	.42	.33**	.41	.42	.02	.37	.43	.11**	.13	.12	.01
35	.17	.19	.04**	.10	.20	.25**	.19	.20	.04	.14	.21	.17**	.10	.09	.03
36	.18	.19	.03*	.12	.20	.21**	.18	.20	.06*	.16	.21	.12**	.25	.26	.00
37	.14	.15	.03*	.09	.16	.19**	.14	.16	.05*	.12	.16	.11**	.16	.17	.01
38	.32	.35	.08**	.26	.37	.23**	.33	.36	.07**	.32	.37	.10**	.26	.23	.02
39	.20	.21	.03*	.17	.22	.13**	.19	.22	.07**	.19	.22	.07**	.17	.16	.02
40	.27	.28	.01	.24	.29	.11**	.25	.29	.08**	.27	.29	.05**	.23	.26	.08

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

Note. M_F = Mean for Female Test-Takers; M_M = Mean for Male Test-Takers; M_B = Mean for Black/African American Test-Takers; M_W = Mean for White Test-Takers; M_A = Mean for Asian Test-Takers; d = Cohen's d . Underlined the effect sizes for the items for which the minority subgroups scored higher than their majority subgroups.

The general pattern of results was consistent among the stratified and unstratified samples, with most of the effect sizes being small to moderate. Large effect sizes were observed for AECP (T1-02 for A/W; T2-14 for A/W; T2-24, T2-27, and T2-29 for B/W). As discussed earlier, the results for the AECP are to be interpreted with caution due to small sample sizes. Large effect sizes were also observed for ANG (T2-26) and for AFRES (T2-25 for B/W; T2-29 for BH/BnH). Most of the

items with moderate effect sizes across the accession sources were observed for B/W comparisons. There were some items for which minority test-takers scored either the same or higher than their majority counterparts (e.g., OTS-CIV T1-19 for F/M; OTS-AD T2-01 for A/W).

Table 63 displays the IRT DIF_{MH} results and Table 64 displays the DIF_{Lord} results for the same subgroups. Seven items on T1 and 10 on T2 were flagged as having either moderate or large DIF for B/W. There was no convergence in DIF among the rest of the subgroups.

Table 63. TR Unstratified Sample DIF_{MH}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.49	.07	.51	.22	.28	.04	.45	.59	.41	.06
02	.23	.35	.32	.01	.01	.24	1.04	.27	.56	.04
03	.26	.33	.18	.15	.01	.32	.65	.42	.14	.04
04	.04	.84	.09	.28	.81	.03	.76	.44	.13	.08
05	.16	.92	.19	.29	.14	.18	1.05	.47	.12	.46
06	.16	.35	.53	.10	.56	.05	.62	.72	.43	.10
07	.24	.11	.11	.27	.04	.26	.89	.74	.28	.01
08	.19	.01	.30	.03	.01	.07	.92	.58	.07	.04
09	.59	.07	.05	.07	.08	.49	.61	.26	.05	.12
10	.23	.56	.35	.10	.07	.19	.35	.12	.54	.74
11	.36	.23	.03	.14	.10	.26	.41	.24	.09	.08
12	.23	.34	.21	.11	.50	.23	.21	.00	.18	.23
13	.63	.36	.23	.13	.08	.07	.63	.47	.31	.21
14	.02	.19	.22	.03	.04	.12	.23	.29	.18	.33
15	.19	.90	.17	.08	.13	.34	.52	.05	.07	.10
16	.24	.74	.02	.19	.06	.33	.85	.02	.30	.10
17	.09	1.24	.17	.05	.19	.11	1.07	.03	.06	.29
18	.11	1.51	.16	.08	.20	.01	.95	.39	.10	.00
19	.14	1.33	.04	.08	.01	.15	.91	.03	.23	.31
20	.04	1.62	.03	.03	.38	.16	1.11	.01	.00	.16
21	.21	1.38	.10	.06	.38	.30	1.13	.10	.04	.18
22	.25	1.67	.11	.23	.11	.20	.99	.10	.25	.06
23	.18	1.70	.07	.10	.07	.12	1.36	.13	.14	.23
24	.11	1.41	.00	.06	.23	.40	1.20	.21	.13	.03
25	.37	1.47	.04	.25	.27	.09	1.12	.09	.17	.12
26	.08	1.48	.09	.04	.30	.30	1.18	.06	.21	.00
27	.15	1.50	.21	.02	.03	.10	1.29	.17	.12	.00
28	.24	1.77	.15	.14	.19	.06	.88	.14	.01	.36
29	.21	.90	.06	.06	.17	.03	.78	.18	.01	.20
30	.13	1.05	.30	.15	.17	.08	.15	.03	.02	.13
31	.17	.82	.22	.21	.03	.12	.13	.18	.09	.51
32	.16	.62	.12	.11	.00	.20	.28	.06	.01	.00
33	.16	.37	.10	.05	.53	.36	.64	.02	.11	.55
34	.02	.06	.13	.15	.00	.17	.13	.05	.04	.02
35	.33	.23	.32	.10	.30	.35	.94	.01	.31	.05
36	.43	.46	.20	.03	.41	.30	.96	.20	.03	.05
37	.36	.87	.10	.19	.16	.30	.88	.10	.03	.20
38	.15	.65	.10	.15	.35	.22	.44	.24	.14	.00
39	.37	.67	.14	.14	.01	.07	.85	.28	.07	.27
40	.12	.11	.06	.00	.07	.15	.84	.33	.09	.30

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Table 64. TR Unstratified Sample DIF_{Lord}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.35	.83	.63	.59	.29	.19	1.00	.66	.74	.13
02	.06	1.09	.46	.40	.02	.38	1.55	.28	.84	.10
03	.12	.95	.25	.40	.04	.19	1.23	.50	.51	.07
04	.17	1.41	.18	.49	.84	.14	1.21	.45	.35	.11
05	.28	1.47	.28	.50	.12	.27	1.40	.45	.31	.44
06	.03	1.14	.72	.32	.61	.09	.80	.49	.48	.01
07	.12	.62	.04	.45	.05	.26	.87	.50	.24	.21
08	.10	.27	.28	.05	.09	.03	1.23	.56	.11	.04
09	.47	.29	.11	.13	.12	.38	1.08	.35	.30	.05
10	.11	.96	.39	.24	.13	.11	.46	.01	.61	.77
11	.22	.83	.07	.40	.14	.15	.62	.22	.29	.07
12	.14	.08	.20	.08	.41	.13	.47	.03	.32	.22
13	.52	.27	.16	.03	.06	.10	.60	.33	.32	.19
14	.08	.37	.18	.04	.11	.10	.20	.13	.11	.21
15	.14	.38	.10	.02	.04	.28	.10	.10	.11	.09
16	.17	.13	.01	.16	.03	.29	.76	.12	.40	.05
17	.07	.66	.22	.10	.25	.13	1.01	.20	.31	.18
18	.04	.94	.23	.32	.27	.06	.90	.12	.23	.04
19	.14	.76	.07	.27	.10	.14	.81	.16	.41	.21
20	.05	1.08	.16	.32	.40	.16	1.02	.15	.27	.10
21	.08	.82	.02	.33	.37	.29	1.15	.25	.42	.04
22	.28	1.16	.20	.53	.05	.19	.93	.04	.45	.08
23	.04	1.31	.18	.52	.12	.16	1.42	.27	.26	.07
24	.02	1.00	.11	.41	.09	.19	1.30	.34	.45	.07
25	.12	1.05	.11	.55	.20	.01	1.24	.25	.46	.11
26	.13	1.05	.20	.34	.26	.30	1.26	.21	.42	.04
27	.02	1.15	.29	.35	.10	.04	1.50	.34	.45	.01
28	.05	1.45	.28	.51	.22	.03	1.09	.30	.28	.24
29	.16	.48	.17	.14	.24	.07	1.14	.30	.32	.20
30	.08	.62	.34	.26	.22	.03	.34	.14	.19	.04
31	.11	.51	.01	.43	.14	.02	.44	.23	.21	.34
32	.11	.31	.05	.16	.13	.09	.46	.03	.25	.05
33	.12	.17	.07	.25	.60	.14	.47	.06	.29	.37
34	.16	.41	.09	.22	.14	.06	.12	.05	.09	.14
35	.29	.16	.09	.01	.07	.21	.14	.11	.46	.04
36	.38	.26	.27	.06	.20	.29	.60	.26	.06	.01
37	.46	.94	.02	.30	.33	.26	.60	.25	.08	.27
38	.35	1.10	.12	.45	.47	.08	.74	.27	.13	.10
39	.51	.96	.18	.43	.11	.26	1.35	.34	.31	.20
40	.14	.93	.11	.55	.16	.40	1.50	.37	.50	.44

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

3.7.9 Conclusion

Overall, the TR subtest items seem to be of below average difficulty and of moderate-to-high discriminability. This subtest is speeded by design. There is some evidence to suggest that overt guessing occurred when examining the longstring option endorsement across the last five items. The removal of any one of the items would have a small effect on the internal consistency of the subtest. Internal consistency is fairly high which may suggest redundancy. Five distractors on T1 and five on T2 were chosen more frequently than the correct response, especially towards the end of the test. Many distractors were chosen infrequently. Common items analysis showed small differences in item difficulty between Forms T1 and T2. Most items showed small-to-moderate (no large) Cohen's *d* across the subgroups. Similarly, most items showed small-to-moderate DIF across the subgroups, with a few items showing moderate-to-large DIF for B/W comparisons.

3.8 Instrument Comprehension (IC)

IC is the eighth cognitive subtest and is one of the five subtests that make up the Classification portion of the AFOQT. It consists of 25 items with four multiple choice options. Test-takers are given 5 minutes to complete the test, for an average of 12 seconds per item. Recall that IC in Form S had 20 items, 5 less than in Form T. Therefore, when making comparisons between the two forms, we looked at general trends rather than at absolute values.

3.8.1 Descriptives

Table 65 summarizes the means, standard deviations, skewness, and kurtosis for the unstratified sample across Forms T1 and T2. The item means ranged from .32 to .75 with an average of .58 on T1 and from .33 to .77 with an average of .58 on T2. The item means generally decrease throughout the test. Skewness tended to switch from negative to positive throughout the test, with the values remaining within the acceptable range.

Table 65. IC Unstratified Sample Descriptives

Item #	T1				T2			
	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
01	.73	.44	-1.05	-.89	.76	.43	-1.22	-.52
02	.73	.44	-1.03	-.94	.76	.43	-1.22	-.52
03	.74	.44	-1.07	-.85	.77	.42	-1.28	-.36
04	.69	.46	-.80	-1.36	.74	.44	-1.12	-.74
05	.68	.46	-.79	-1.37	.74	.44	-1.08	-.84
06	.70	.46	-.89	-1.20	.72	.45	-.96	-1.08
07	.67	.47	-.70	-1.51	.67	.47	-.70	-1.51
08	.68	.47	-.76	-1.42	.63	.48	-.55	-1.70
09	.75	.44	-1.12	-.73	.74	.44	-1.07	-.85
10	.75	.43	-1.15	-.67	.68	.47	-.78	-1.40
11	.65	.48	-.64	-1.60	.74	.44	-1.12	-.74
12	.65	.48	-.63	-1.61	.54	.50	-.18	-1.97
13	.70	.46	-.85	-1.28	.65	.48	-.64	-1.59
14	.60	.49	-.43	-1.82	.61	.49	-.45	-1.80
15	.60	.49	-.43	-1.82	.56	.50	-.26	-1.93
16	.50	.50	.00	-2.00	.50	.50	.00	-2.00
17	.52	.50	-.07	-1.99	.49	.50	.05	-2.00
18	.42	.49	.32	-1.90	.52	.50	-.09	-1.99
19	.54	.50	-.15	-1.98	.41	.49	.36	-1.87
20	.41	.49	.35	-1.87	.43	.50	.28	-1.92
21	.37	.48	.55	-1.70	.41	.49	.38	-1.85
22	.32	.47	.76	-1.43	.33	.47	.74	-1.45
23	.45	.50	.22	-1.95	.38	.49	.49	-1.76
24	.41	.49	.36	-1.87	.44	.50	.25	-1.94
25	.35	.48	.63	-1.61	.34	.47	.69	-1.53

Note. *M* = Mean; *SD* = Standard Deviation. $N_{T1} = 36,070$; $N_{T2} = 32,930$.

Results for the stratified samples were similar to the unstratified sample in terms of decreasing means and skewness switching from negative to positive throughout the test. The USAFA had the highest mean among the accession sources, with the T1 mean being equal to .63 and T2 mean being equal to .62. Carretta et al. (2016), Aguilar (2017), and Morath et al. (2008) do not report item-level central tendency statistics.

3.8.2 Difficulty

Table 66 summarizes the *p*-values and 2PL *b* parameters for the unstratified sample across Forms T1 and T2. The *p*-values ranged widely on both Forms T1 and T2, but in general, they tended to decrease throughout the test. None of the items showed ceiling or floor effects. The 2PL *b* parameters tended to switch from negative to positive throughout the test, while staying within the acceptable range.

Table 66. IC Unstratified Sample Difficulty

Item #	T1		T2	
	<i>p</i> -values	2PL <i>b</i>	<i>p</i> -values	2PL <i>b</i>
01	.73	-1.00	.76	-.94
02	.73	-.81	.76	-.96
03	.74	-.84	.77	-1.40
04	.69	-.61	.74	-1.24
05	.68	-.62	.74	-.83
06	.70	-.73	.72	-.83
07	.67	-.57	.67	-.59
08	.68	-.62	.63	-.46
09	.75	-.87	.74	-.82
10	.75	-.91	.68	-.69
11	.65	-.54	.74	-.95
12	.65	-.59	.54	-.21
13	.70	-.95	.65	-.95
14	.60	-.43	.61	-.39
15	.60	-.38	.56	-.25
16	.50	-.01	.50	-.02
17	.52	-.08	.49	.02
18	.42	.25	.52	-.10
19	.54	-.14	.41	.27
20	.41	.39	.43	.23
21	.37	.51	.41	.42
22	.32	.67	.33	.64
23	.45	.18	.38	.67
24	.41	.44	.44	.21
25	.35	1.10	.34	1.19

Note. *p*-value = Difficulty; 2PL *b* = IRT Difficulty Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing *p*-values and 2PL *b* parameters switching from negative to positive throughout the test. From the CTT perspective, the USAFA scored higher on average than any other accession source, with T1 averaging at .63 and T2 averaging at .62. There was one item on T1 and one item on T2 that reached a ceiling effect for USAFA. No items reached a floor effect across the stratification samples. Notably, all items with the highest *p*-values were towards the beginning of the test and all items with the lowest *p*-values were towards the end of the test across the accession sources. From the IRT perspective, all items across the accession sources were within the acceptable range.

The *p*-values reported by Carretta et al. (2016) and Aguilar (2017) for Form T were as follows: T1 *p*-value = .61 and T2 *p*-value = .60. The current *p*-values are lower: T1 *p*-value = .58 and T2 *p*-value = .58). These slight differences may be due to the present study utilizing a larger sample. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seems to be more difficult. The mean *p*-values for Form S1 and S2 were .64 and .61, respectively. However, remember that the Form T IC subtest included five more items than its Form S counterpart and the administration time got shortened from 6 to 5 minutes. Therefore, comparing the two subtests may not be as clear cut.

3.8.3 Discriminability

Table 67 summarizes the ITC's and 2PL a parameters for the unstratified sample across Forms T1 and T2. ITCs ranged from .33 to .66 with an average of .57 for T1 and from .32 to .65 with an average of .55 on T2. Two items on T1 and three items on T2 had unacceptable 2PL a parameter.

Table 67. IC Unstratified Sample Discriminability

Item #	T1		T2	
	ITC	2PL a	ITC	2PL a
01	.50	1.36	.58	2.10
02	.61	2.10	.57	1.97
03	.60	2.07	.42	1.05
04	.66	2.50	.43	1.07
05	.65	2.30	.61	2.22
06	.61	1.98	.56	1.67
07	.64	2.11	.62	2.02
08	.63	2.08	.64	2.13
09	.60	2.10	.61	2.24
10	.59	1.96	.58	1.67
11	.62	1.91	.56	1.71
12	.56	1.52	.51	1.18
13	.46	1.09	.38	.75
14	.54	1.38	.64	2.04
15	.63	1.87	.62	1.80
16	.60	1.66	.64	1.94
17	.65	2.08	.61	1.75
18	.63	1.98	.62	1.82
19	.64	1.93	.65	2.11
20	.48	1.11	.60	1.70
21	.54	1.48	.48	1.12
22	.54	1.57	.56	1.67
23	.61	1.77	.40	.83
24	.43	.96	.60	1.67
25	.33	.61	.32	.61

Note. ITC = Item-Total Correlation; 2PL a = IRT Discriminability Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample, such that there was a wide range of ITC's among the items and the majority of the items had acceptable-to-high 2PL a parameters.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and Aguilar (2017) for Form T. When comparing the present findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seemed to be equally discriminable (T1 = .57 and T2 = .55 vs. S1 = .52 and S2 = .52).

3.8.4 Internal Consistency if Item is Deleted

Table 68 summarizes Cronbach's alphas if an item is deleted for the unstratified sample across Forms T1 and T2. For reference, the IC subtest-level alphas for T1 and for T2 were .91 and .91,

respectively. The item-level alphas were high, ranging from .91 to .92 on T1 and from .90 to .91 on T2. The removal of any one of the items would have a small effect on the internal consistency of the subtest.

Table 68. IC Unstratified Sample Internal Consistency if Item is Deleted

Item #	T1 α	T2 α
01	.91	.90
02	.91	.90
03	.91	.90
04	.91	.90
05	.91	.90
06	.91	.90
07	.91	.90
08	.91	.90
09	.91	.90
10	.91	.90
11	.91	.90
12	.91	.90
13	.91	.91
14	.91	.90
15	.91	.90
16	.91	.90
17	.91	.90
18	.91	.90
19	.91	.90
20	.91	.90
21	.91	.90
22	.91	.90
23	.91	.91
24	.91	.90
25	.92	.91

Note. α = Cronbach's Alpha.

The results for the stratified samples were similar to the unstratified sample, such that the alphas were all in the .90 range and no one item was indicated as dramatically improving internal consistency if deleted.

Although the reports by Aguilar (2017), Carretta et al. (2016), and Morath et al. (2008) did not report the values for alpha if items were deleted, the Cronbach's alpha they reported for the subtests were similar to the alphas observed in this research.

3.8.5 Omission Rates and Overt Guessing

Table 69 summarizes the omission rates in a form of missing values for each item and for a grouping of five items for Forms T1 and T2. The longstring omission rate across the last five items was 12.00% for T1 and 12.54% for T2. In general, the omission rates increased throughout the test. For T1; the average omission rate on the first five items was .22%, whereas the average on

the last five items was 16.55%. For T2, these averages went from .20% to 17.14%. The longstring overt guessing for the last five items was 2.81% for T1 and 2.26% for T2.

Table 69. IC Unstratified Sample Omission Rates

Item #	T1		T2	
	For Each Item	Across Five Items	For Each Item	Across Five Items
01	.26%	.22%	.09%	.20%
02	.09%		.18%	
03	.21%		.26%	
04	.26%		.25%	
05	.26%		.20%	
06	.20%	.38%	.15%	.40%
07	.29%		.28%	
08	.32%		.34%	
09	.47%		.49%	
10	.61%		.75%	
11	.93%	2.27%	.95%	2.33%
12	1.26%		1.41%	
13	2.30%		2.15%	
14	2.84%		3.15%	
15	4.03%		4.02%	
16	5.15%	8.05%	5.14%	8.26%
17	6.56%		6.77%	
18	7.90%		8.09%	
19	9.47%		9.88%	
20	11.14%		11.42%	
21	13.32%	16.55%	13.95%	17.14%
22	14.97%		15.69%	
23	16.55%		17.09%	
24	17.93%		18.65%	
25	20.00%		20.34%	

Note. For Each Item = Percent of Missing Values in Each Item; Across Five Items = Percent of Missing Values Across a Grouping of five Items.

The results for the stratified samples were similar to those for the unstratified sample, such that the omission rates tended to increase throughout the test. The ANG had the lowest omission rate in the grouping of the last five items (T1: 13.81%; T2: 14.56%).

In general, the current findings are consistent with the results reported by Morath et al. (2008) for Form s and by Carretta et al. (2016) for Form T for the unstratified sample. These reports did not included the omission and guessing rates for the stratified samples. Aguilar (2017) did not report omission rates or overt guessing.

3.8.6 Distractor Analysis

When examining the response frequencies, the correct response was chosen more frequently than a distractor for all items on both Forms T1 and T2.

The pattern of results for the stratified samples was similar to the unstratified sample, such that all accession source chose the keyed response more often than a distractor.

In general, the current findings are consistent with the results reported by Carretta et al. (2016) for Form T. Aguilar (2017) did not report the distractor analysis. When comparing the present results against the results for Form S reported by Morath et al. (2008) for the unstratified sample, all items on Form S seemed to have had no items for which the distractor was chosen more often. The authors mentioned that when examining stratified response frequencies, the correct response was the most frequently chosen response on all but items S2-18 and S2-20 for the AECF accession source. Given qualitative differences between Form S and Form T, the comparison is not as clear.

3.8.7 Common Items

Thirteen items were common between Forms T1 and T2. They are displayed in Table 70. The table also displays the results of the independent samples *t*-tests comparing the *p*-values for these common items. As can be seen, five *t*-tests were statistically significant, but the effect sizes (Cohen's *d*) were trivial, ranging from .01 to .13. The rest of the *t*-tests were trivial as well, with Cohen's *d* being equal to .01. Note that the *t*-tests were run for unstratified sample only. A difference in *p*-values was also examined for the stratified samples only. The differences between the *p*-values were small, ranging from |.00| to |.02| (see Attachment 1).

Table 70. IC Unstratified Sample Common Items Difficulty

T1 Item #	T2 Item #	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	Sig.	ES
T1-08	T2-07	.68	.47	.67	.47	.00	.03
T1-14	T2-15	.61	.49	.56	.50	.00	.08
T1-15	T2-14	.60	.49	.61	.49	.28	.01
T1-16	T2-17	.50	.50	.49	.50	.00	.02
T1-17	T2-16	.52	.50	.50	.50	.00	.03
T1-18	T2-19	.42	.49	.41	.49	.03	.02
T1-19	T2-18	.54	.50	.52	.50	.00	.03
T1-20	T2-21	.41	.49	.41	.49	.09	.01
T1-21	T2-20	.37	.48	.43	.50	.00	.13
T1-22	T2-22	.32	.47	.33	.47	.33	.01
T1-23	T2-24	.45	.50	.44	.50	.07	.01
T1-24	T2-23	.41	.49	.38	.49	.00	.06
T1-25	T2-25	.35	.48	.34	.47	.00	.03

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; Sig. = Significance Levels; ES = Effect Size.

The only technical report that included an examination of the common items was by Morath et al. (2008) for Form S. Similar to the current findings, the authors did not detect sizable differences among the common items on Form S.

3.8.8 Subgroup Differences

Table 71 summarizes the CTT effect sizes (Cohen's *d*) for the mean score difference among the gender, racial, and ethnic subgroups across Forms T1 and T2.

Regarding comparisons between the Female (F) and Male (M) subgroups, most of the effect sizes were moderate. There was one item on T1 (T1-25) and four items on T2 (T2-03, T2-13, T2-23, and T2-25) whose effect sizes were small. On T1, the effect sizes ranged from .30 to .72, with M scoring higher than F on all 25 items. On T2, the effect sizes ranged from .26 to .70, with M scoring higher than F on 21 items.

For the Black (B) and White (W) subgroups, most of the effect sizes were moderate. There were two items with small and one with large effect sizes on T1. There were three items with small and two with large effect sizes on T2. On T1, the effect sizes ranged from .31 to .80, with W scoring higher than B on all 25 items. On T2, the effect sizes ranged from .31 to .83, with W scoring higher than B on all 25 items.

For comparisons between the Asian (A) and White (W) subgroups, all effect sizes were small. On T1, the effect sizes ranged from .12 to .34, with W scoring higher than A on all 25 items. On T2, the effect sizes ranged from .08 to .31, with W scoring higher than A on 24 items.

Regarding the White Hispanic (WH) and White Non-Hispanic (WnH) subgroups, all effect sizes were small. On T1, the effect sizes ranged from .10 to .26, with WnH scoring higher than WH on 25 items. On T2, effect sizes ranged from .08 to .23, with WnH scoring higher than WH on 25 items.

For comparisons between the Black Hispanic (BH) and Black Non-Hispanic (BnH) subgroups, all effect sizes were small. On T1, the effect sizes ranged from .00 to .16, with BnH scoring higher than BH on one item. On T2, effect sizes ranged from .01 to .20, with BnH scoring higher than BH on one item.

Table 71. IC Unstratified Sample Mean Score Subgroup Differences - Cohen's *d*

Item #	T1														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.51	.81	.71**	.49	.79	.71**	.66	.78	.29**	.71	.79	.20**	.59	.51	.16**
02	.52	.80	.66**	.47	.78	.73**	.67	.78	.27**	.70	.79	.22**	.55	.49	.11*
03	.57	.79	.51**	.52	.79	.63**	.69	.78	.22**	.71	.79	.20**	.56	.53	.06
04	.46	.77	.70**	.43	.74	.72**	.61	.74	.29**	.66	.75	.20**	.50	.45	.09†
05	.46	.76	.68**	.40	.75	.80**	.60	.75	.34**	.66	.76	.21**	.47	.42	.09†
06	.51	.78	.61**	.44	.76	.74**	.65	.76	.25**	.66	.77	.26**	.53	.46	.14**
07	.45	.74	.64**	.39	.72	.73**	.62	.72	.22**	.65	.73	.18**	.48	.42	.13**
08	.47	.75	.61**	.43	.73	.67**	.62	.73	.25**	.65	.74	.20**	.50	.45	.10*
09	.57	.81	.56**	.53	.80	.65**	.69	.79	.26**	.72	.80	.21**	.59	.54	.10*
10	.57	.81	.58**	.53	.80	.65**	.69	.80	.25**	.71	.81	.23**	.59	.55	.06
11	.41	.74	.72**	.38	.71	.74**	.57	.71	.30**	.62	.72	.21**	.45	.40	.11*
12	.47	.71	.53**	.44	.70	.57**	.60	.69	.20**	.62	.70	.18**	.46	.45	.01
13	.50	.76	.58**	.52	.74	.48**	.65	.74	.19**	.66	.74	.18**	.53	.54	.00
14	.42	.67	.52**	.37	.66	.60**	.57	.65	.18**	.58	.66	.17**	.42	.39	.07
15	.39	.68	.63**	.32	.67	.74**	.53	.67	.29**	.56	.68	.24**	.39	.35	.07
16	.29	.57	.58**	.27	.55	.57**	.43	.55	.25**	.45	.56	.22**	.32	.30	.05
17	.30	.60	.60**	.25	.58	.67**	.43	.58	.30**	.46	.59	.26**	.32	.27	.10*
18	.21	.50	.61**	.19	.48	.59**	.37	.47	.21**	.37	.48	.22**	.25	.20	.11*
19	.32	.61	.61**	.29	.60	.64**	.45	.59	.29**	.49	.61	.24**	.32	.30	.04
20	.26	.47	.43**	.25	.45	.42**	.37	.45	.17**	.38	.46	.16**	.26	.26	.01
21	.18	.43	.54**	.18	.41	.50**	.30	.41	.22**	.32	.42	.21**	.21	.19	.05
22	.15	.39	.53**	.15	.37	.46**	.26	.36	.21**	.27	.37	.21**	.18	.17	.03
23	.23	.52	.60**	.22	.50	.58**	.39	.50	.22**	.40	.51	.21**	.26	.23	.06
24	.26	.47	.43**	.27	.45	.35**	.34	.45	.22**	.38	.45	.15**	.30	.28	.04
25	.25	.39	.30**	.23	.38	.31**	.32	.38	.12**	.33	.38	.10**	.24	.24	.01
Item #	T2														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.58	.82	.59**	.52	.81	.72**	.72	.81	.22**	.76	.81	.13**	.63	.54	.19**
02	.58	.82	.58**	.55	.81	.64**	.71	.81	.25**	.74	.81	.19**	.63	.56	.14**
03	.66	.81	.36**	.60	.80	.50**	.76	.80	.10**	.76	.81	.12**	.68	.61	.15**
04	.59	.80	.51**	.56	.79	.54**	.71	.78	.16**	.73	.79	.14**	.60	.58	.06
05	.52	.81	.70**	.46	.80	.83**	.68	.80	.30**	.74	.80	.17**	.56	.47	.18**
06	.55	.77	.51**	.50	.77	.61**	.68	.76	.19**	.71	.77	.14**	.57	.51	.12*
07	.45	.74	.64**	.42	.72	.67**	.60	.72	.26**	.64	.73	.18**	.51	.44	.15**
08	.41	.71	.65**	.37	.69	.71**	.55	.69	.31**	.61	.70	.20**	.47	.39	.16**
09	.53	.81	.68**	.45	.80	.83**	.69	.80	.25**	.72	.80	.20**	.58	.48	.20**
10	.48	.75	.59**	.47	.73	.59**	.64	.73	.19**	.67	.73	.14**	.56	.48	.17**
11	.57	.80	.54**	.51	.79	.68**	.72	.79	.18**	.73	.80	.15**	.60	.52	.16**
12	.35	.61	.55**	.33	.59	.54**	.48	.59	.23**	.51	.60	.18**	.43	.34	.17**
13	.54	.69	.32**	.51	.68	.37**	.64	.68	.08**	.64	.68	.10**	.54	.53	.03
14	.38	.69	.66**	.32	.68	.76**	.53	.67	.30**	.58	.68	.22**	.40	.34	.13**
15	.36	.64	.56**	.30	.62	.68**	.53	.62	.19**	.54	.63	.19**	.41	.31	.20**
16	.29	.58	.60**	.27	.55	.59**	.43	.55	.25**	.47	.56	.17**	.34	.29	.13**
17	.28	.56	.58**	.27	.54	.55**	.42	.54	.25**	.43	.55	.23**	.34	.28	.13**
18	.31	.60	.59**	.28	.58	.61**	.46	.58	.25**	.48	.59	.21**	.36	.29	.14**
19	.19	.49	.62**	.19	.47	.56**	.35	.46	.22**	.37	.47	.19**	.22	.20	.03
20	.21	.51	.62**	.21	.49	.57**	.34	.49	.29**	.39	.50	.22**	.25	.22	.08†
21	.26	.46	.42**	.24	.45	.43**	.36	.44	.18**	.39	.45	.13**	.29	.25	.09†
22	.14	.39	.54**	.14	.37	.49**	.26	.37	.23**	.27	.38	.22**	.15	.16	.02
23	.25	.43	.36**	.25	.42	.34**	.32	.41	.19**	.34	.42	.17**	.28	.25	.07
24	.23	.51	.58**	.21	.49	.58**	.39	.49	.19**	.39	.50	.23**	.23	.23	.01
25	.25	.37	.26**	.22	.37	.31**	.31	.36	.10**	.33	.37	.08**	.27	.23	.10*

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

Note. M_F = Mean for Female Test-Takers; M_M = Mean for Male Test-Takers; M_B = Mean for Black/African American Test-Takers; M_W = Mean for White Test-Takers; M_A = Mean for Asian Test-Takers; d = Cohen's d . Underlined the effect sizes for the items for which the minority subgroups scored higher than their majority subgroups.

The general pattern of results was consistent among the stratified and unstratified samples, such that for F/M and for B/W comparisons, the majority of the items had either moderate or large effect sizes. The effect sizes for the A/W and for the ethnic subgroups were primarily small with a few items showing moderate effect sizes. The accession source with the highest number of items with large effect sizes was AFRES on both T1 and T2. T1-05 and T1-11 showed a consistent pattern of large effect sizes for B/W across most of the accession sources. Similar pattern was observed for T2-05 and T2-09.

Table 72 displays the IRT DIF_{MH} results and Table 73 displays the DIF_{Lord} results for the same subgroups. No items were flagged simultaneously by the two methods on T1. T2-05 was flagged as showing moderate effect size for B/W.

Table 72. IC Unstratified Sample DIF_{MH}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	1.35	1.02	.49	.17	.59	.02	.46	.12	.35	.36
02	.72	.59	.18	.09	.08	.03	.14	.35	.10	.05
03	.00	.03	.16	.09	.22	.44	.27	.35	.07	.18
04	.79	.30	.17	.26	.15	.38	.39	.06	.05	.20
05	.72	.90	.58	.02	.06	.53	1.06	.60	.14	.11
06	.42	.66	.07	.36	.42	.50	.08	.09	.18	.07
07	.52	.52	.24	.27	.13	.07	.13	.31	.06	.19
08	.35	.19	.04	.10	.11	.02	.29	.57	.12	.03
09	.19	.07	.17	.05	.24	.41	.99	.25	.13	.03
10	.25	.18	.10	.22	.10	.01	.13	.11	.27	.30
11	.96	.68	.33	.03	.18	.08	.37	.18	.16	.08
12	.03	.12	.23	.03	.16	.17	.11	.26	.15	.32
13	.79	.08	.03	.12	.23	.41	.08	.32	.08	.24
14	.09	.38	.28	.06	.12	.09	.73	.48	.18	.13
15	.23	.70	.23	.14	.02	.51	.39	.23	.01	.30
16	.01	.19	.01	.08	.22	.29	.19	.13	.22	.26
17	.41	.16	.20	.14	.18	.23	.44	.17	.24	.03
18	.06	.04	.43	.07	.31	.25	.04	.14	.07	.12
19	.13	.02	.19	.05	.29	.19	.47	.21	.15	.81
20	.29	.30	.20	.06	.22	.17	.10	.42	.11	.12
21	.09	.01	.05	.08	.14	.44	.27	.07	.25	.14
22	.20	.32	.16	.12	.17	.21	.22	.11	.21	.80
23	.03	.08	.35	.09	.00	.39	.22	.20	.17	.11
24	.27	.41	.22	.07	.02	.16	.05	.23	.19	.69
25	.21	.11	.04	.09	.17	.51	.13	.16	.13	.13

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Table 73. IC Unstratified Sample DIF_{Lord}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.74	.42	.33	.02	.57	.39	.61	.13	.18	.45
02	.44	.57	.22	.16	.05	.36	.21	.30	.16	.18
03	.24	.06	.08	.00	.16	.86	.48	.65	.26	.18
04	.67	.46	.31	.01	.09	.11	.32	.24	.12	.26
05	.53	.93	.60	.10	.03	.91	1.16	.56	.01	.30
06	.20	.58	.07	.36	.31	.14	.05	.10	.11	.05
07	.31	.54	.11	.13	.11	.50	.35	.29	.05	.33
08	.16	.24	.06	.02	.11	.51	.55	.56	.22	.18
09	.03	.14	.22	.14	.21	.80	1.12	.26	.23	.20
10	.12	.16	.17	.27	.04	.23	.12	.11	.18	.35
11	.74	.60	.35	.05	.14	.09	.41	.15	.06	.17
12	.37	.35	.28	.10	.15	.06	.34	.03	.04	.31
13	.06	.79	.29	.13	.33	1.34	1.33	.81	.40	.42
14	.40	.12	.41	.18	.06	.52	.90	.46	.28	.05
15	.19	.66	.27	.18	.03	.03	.49	.18	.08	.39
16	.09	.19	.01	.07	.20	.24	.06	.17	.04	.13
17	.01	.50	.29	.24	.15	.14	.14	.16	.28	.01
18	.32	.27	.22	.05	.26	.17	.14	.14	.16	.15
19	.08	.24	.27	.14	.29	.55	.18	.00	.08	.65
20	.96	1.01	.49	.30	.28	.50	.17	.42	.17	.09
21	.07	.30	.10	.01	.11	.79	.82	.28	.35	.26
22	.05	.40	.15	.08	.16	.27	.05	.13	.26	.72
23	.17	.04	.19	.01	.00	1.23	1.40	.17	.05	.13
24	.88	1.51	.13	.38	.14	.19	.19	.19	.24	.62
25	1.75	1.75	.75	.67	.25	1.81	1.59	.75	.54	.03

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Aguilar (2017) and Carretta et al. (2016) did not include either CTT or IRT subgroup difference analyses in their report. Morath et al. (2008) included CTT Cohen's *d* results for Form S. The authors observed large effect sizes for F/M and for B/W.

3.8.9 Conclusion

Overall, the items in the IC subtest seem to be of below average difficulty and of high discriminability. Item omission rates and longstring option endorsement across the last five items suggest speededness and overt guessing. The removal of any one of the items would have a small effect on the internal consistency of the IC subtest. Internal consistency is fairly high which may suggest redundancy. No distractors were chosen more frequently than the correct response. No distractors on T1 and five on T2 were chosen infrequently. Common items analysis showed small differences in item difficulty between Forms T1 and T2. Most items showed moderate Cohen's *d*, especially for F/M and B/W subgroup comparisons, with a few items showing large Cohen's *d* for B/W comparisons. Similarly, most items showed small-to-moderate DIF across the subgroups.

3.9 Block Counting

Block Counting (BC) is the ninth cognitive subtest and is one of the five subtests that make up the Classification part of the AFOQT. It consists of 30 items with five multiple choice options. Test-takers are allotted 4.5 minutes to complete the test, for an average of 9 seconds per item. Note that unlike the rest of the subtests, BC presented test-takers with five blocks on T1 and five blocks on T2, each corresponding to a set of five items. Thus, each group of five items belonged to one block. Therefore, when evaluating the psychometric of this subtest, we looked at the general pattern across the groupings of five items.

3.9.1 Descriptives

Table 74 summarizes the means, standard deviations, skewness, and kurtosis for the unstratified sample across Forms T1 and T2. The item means ranged from .08 to .86 with an average of .51 on T1 and from .17 to .91 with an average of .54 on T2. The means generally decreased throughout the test, but this trend was less pronounced than in the other subtests. The means for the first five items exceeded .45, whereas the means for the last five items fell below .45 across Forms T1 and T2. Skewness tended to switch from negative to positive throughout the test, but the pattern was not as clear as in the other subtests. While the majority of the skewness values were within the acceptable range, T1-27 had an extreme value (3.12).

Table 74. BC Unstratified Sample Descriptives

Item #	T1				T2			
	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
01	.49	.50	.02	-2.00	.86	.35	-2.02	2.09
02	.68	.46	-.80	-1.37	.68	.47	-.79	-1.38
03	.56	.50	-.23	-1.95	.81	.39	-1.58	.49
04	.69	.46	-.83	-1.32	.91	.29	-2.85	6.12
05	.74	.44	-1.09	-.82	.75	.43	-1.17	-.64
06	.70	.46	-.86	-1.26	.73	.44	-1.05	-.90
07	.74	.44	-1.11	-.77	.60	.49	-.40	-1.84
08	.75	.43	-1.14	-.70	.68	.47	-.79	-1.38
09	.84	.37	-1.86	1.45	.84	.36	-1.88	1.54
10	.86	.35	-2.09	2.36	.87	.33	-2.24	3.02
11	.50	.50	-.01	-2.00	.76	.43	-1.22	-.52
12	.19	.40	1.55	.40	.59	.49	-.38	-1.86
13	.27	.44	1.05	-.90	.55	.50	-.22	-1.95
14	.37	.48	.53	-1.72	.90	.30	-2.69	5.26
15	.40	.49	.40	-1.84	.54	.50	-.17	-1.97
16	.74	.44	-1.11	-.76	.46	.50	.15	-1.98
17	.57	.50	-.28	-1.92	.18	.38	1.68	.82
18	.49	.50	.06	-2.00	.25	.43	1.18	-.61
19	.73	.44	-1.03	-.94	.33	.47	.72	-1.48
20	.42	.49	.31	-1.90	.33	.47	.70	-1.50
21	.55	.50	-.19	-1.96	.68	.47	-.76	-1.42
22	.60	.49	-.40	-1.84	.17	.38	1.71	.94
23	.48	.50	.08	-1.99	.31	.46	.82	-1.32
24	.52	.50	-.08	-1.99	.25	.43	1.17	-.62
25	.35	.48	.61	-1.62	.53	.50	-.13	-1.98
26	.18	.39	1.64	.70	.44	.50	.24	-1.94
27	.08	.27	3.12	7.71	.25	.44	1.13	-.72
28	.31	.46	.81	-1.34	.26	.44	1.09	-.80
29	.14	.35	2.03	2.11	.22	.42	1.34	-.21
30	.27	.44	1.04	-.92	.32	.47	.75	-1.44

Note. *M* = Mean; *SD* = Standard Deviation. $N_{T1} = 36,039$ $N_{T2} = 32,904$.

The pattern of results for the stratified samples was similar to the unstratified sample. Even though the item means generally decreased throughout the test for the accession sources, this trend was less pronounced than in the rest of the subtests. No clear pattern was observed with skewness either. The USAFA had a lower average mean than the other accession sources, with T1 averaging at .46 and T2 averaging at .50. Carretta et al. (2016), Aguilar (2017), and Morath et al. (2008) do not report item-level central tendency statistics.

3.9.2 Difficulty

Table 75 summarizes the *p*-values and 2PL *b* parameters for the unstratified sample across Forms T1 and T2. The *p*-values ranged widely on both Forms T1 and T2, with some decrease throughout the test but as pronounced as the rest of the subtests. On T1, two items reached a ceiling effect and four items reached a floor effect. On T2, six items reached a ceiling effect and two items reached a floor effect. Notably, all items with ceiling effects appeared before the items with a floor effect.

There was no clear pattern in the 2PL b parameters either. One item on T1 and six items on T2 had extreme 2PL b values.

Table 75. BC Unstratified Sample Difficulty

Item #	T1		T2	
	p -values	2PL b	p -values	2PL b
01	.49	.02	.86	-2.00
02	.68	-.72	.68	-.79
03	.56	-.22	.81	-2.27
04	.69	-.99	.91	-4.33
05	.74	-1.00	.75	-.84
06	.70	-1.07	.73	-1.29
07	.74	-1.02	.60	-.37
08	.75	-.88	.68	-.55
09	.84	-2.24	.84	-2.47
10	.86	-1.66	.87	-1.79
11	.50	-.01	.76	-2.49
12	.19	1.70	.59	-.33
13	.27	1.19	.55	-.21
14	.37	.73	.90	-4.11
15	.40	.38	.54	-.16
16	.74	-1.47	.46	.18
17	.57	-.30	.18	1.71
18	.49	.06	.25	1.20
19	.73	-1.49	.33	.83
20	.42	.29	.33	.63
21	.55	-.20	.68	-1.24
22	.60	-.44	.17	2.22
23	.48	.09	.31	1.84
24	.52	-.12	.25	1.02
25	.35	.81	.53	-.29
26	.18	1.75	.44	.51
27	.08	2.37	.25	1.48
28	.31	1.29	.26	1.55
29	.14	1.64	.22	1.69
30	.27	1.54	.32	1.91

Note. p -value = Difficulty; 2PL b = IRT Difficulty Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample. Specifically, while p -values tended to decrease, this trend was less pronounced than in the other subtests. All accession sources had at least two items with floor effects and at least one item with a ceiling effect. Similar to the unstratified sample, all items with a ceiling effect appeared before the items with floor effects across the accession sources. From the IRT perspective, the USAFA had the highest number of items with unacceptable 2PL b parameters (four on T1 and 10 on T2).

In general, the current findings are consistent with the results reported by Carretta et al. (2016) and by Aguilar (2017) for Form T. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seems to be more difficult. The mean p -values

for Forms T1 and T2 were .51 and .54, respectively. The mean p -values for S1 and S2 were .61 and .60, respectively.

3.9.3 Discriminability

Table 76 summarizes the ITC's and 2PL a parameters for the unstratified sample across Forms T1 and T2. ITC's ranged from .30 to .55 with an average of .44 for T1 and from .21 to .60 with an average of .40 on T2. Twelve items on T1 and 15 items on T2 had unacceptable 2PL a parameter.

Table 76. BC Unstratified Sample Discriminability

Item #	T1		T2	
	ITC	2PL a	ITC	2PL a
01	.50	1.29	.36	1.07
02	.50	1.51	.45	1.24
03	.53	1.51	.31	.70
04	.40	.98	.21	.56
05	.48	1.42	.55	2.30
06	.38	.91	.38	.91
07	.46	1.39	.51	1.50
08	.54	1.96	.60	2.59
09	.30	.84	.30	.75
10	.41	1.48	.40	1.43
11	.44	.97	.26	.48
12	.37	.99	.52	1.61
13	.41	1.01	.49	1.36
14	.39	.83	.23	.57
15	.53	1.40	.52	1.49
16	.36	.82	.45	1.04
17	.49	1.21	.36	1.08
18	.48	1.13	.41	1.18
19	.36	.74	.42	1.04
20	.55	1.47	.52	1.56
21	.55	1.32	.38	.65
22	.51	1.12	.31	.78
23	.48	1.01	.26	.46
24	.45	.87	.53	1.52
25	.43	.86	.35	.46
26	.37	1.02	.37	.50
27	.32	1.30	.41	.83
28	.35	.68	.43	.76
29	.44	1.47	.43	.85
30	.34	.72	.31	.40

Note. ITC = Item-Total Correlation; 2PL a = IRT Discriminability Parameter.

The results for the stratified samples were similar to the unstratified sample, such that there was a wide range of ITC's among the items and the majority of the items had acceptable 2PL a parameters. The only exception was the USAFA, for which negative 2PL a parameters were observed (one item on T1 and four items on T2). This can suggest that either the model-data

convergence was inadequate or the items do not behave as expected. Because none of the other accession sources had negative 2PL α parameters, it is likely that the items function well and the best explanation for the negative values lies in the inadequate convergence.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and by Aguilar (2017) for Form T. When comparing the present findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seemed to be more discriminable ($T1 = .44$ and $T2 = .40$ vs $S1 = .39$ and $S2 = .38$).

3.9.4 Internal Consistency if Item is Deleted

Table 77 summarizes Cronbach's alphas if an item is deleted for the unstratified sample across Forms T1 and T2. For reference, the BC subtest-level alphas for T1 and for T2 were .85 and .82, respectively. The item-level alphas were high, all .85 on T1 and ranging from .81 to .82 on T2. The removal of any one of the items would have a small effect on the internal consistency of the subtest.

Table 77. BC Unstratified Sample Internal Consistency if Item is Deleted

Item #	T1 α	T2 α
01	.85	.82
02	.85	.81
03	.85	.82
04	.85	.81
05	.85	.81
06	.85	.82
07	.85	.82
08	.85	.82
09	.85	.81
10	.85	.82
11	.85	.82
12	.85	.81
13	.85	.82
14	.85	.82
15	.85	.82
16	.85	.82
17	.85	.81
18	.85	.82
19	.85	.82
20	.85	.82
21	.85	.81
22	.85	.82
23	.85	.82
24	.85	.82
25	.85	.82
26	.85	.82
27	.85	.82
28	.85	.82
29	.85	.81
30	.85	.82

Note. α = Cronbach's Alpha.

The pattern of results for the stratified samples was similar to the unstratified sample, such that most of the alphas were within the acceptable range (in the .80s) and no one item was indicated as dramatically improving internal consistency if deleted. The only exception to that was the USAFA on T2, for which the alphas were in the .70s range. Although the reports by Aguilar (2017), Carretta et al. (2016), and Morath et al. (2008) did not report the values for alpha if items were deleted, the Cronbach's alpha they reported for the subtests were similar to the alphas observed in this research.

3.9.5 Omission Rates and Overt Guessing

Table 78 summarizes the percentages of the omission rates in a form of missing values for each item and for a grouping of five items for Forms T1 and T2. The longstring omission rate across the last five items was 17.55% for T1 and 17.93% for T2. In general, the omission rates increased throughout the test. For T1, the average omission rate on the first five items was .14%, whereas the average on the last five items was 21.65%. For T2, these averages went from .11% to 21.90%. The longstring overt guessing for the last five items was 3.15% for T1 and 3.68% for T2.

Table 78. BC Unstratified Sample Omission Rates

Item #	T1		T2	
	For Each Item	Across Five Items	For Each Item	Across Five Items
01	.10%	.22%	.05%	.17%
02	.12%		.46%	
03	.25%		.13%	
04	.39%		.11%	
05	.24%		.09%	
06	.05%	.14%	.08%	.11%
07	.07%		.07%	
08	.10%		.11%	
09	.20%		.15%	
10	.25%		.16%	
11	.39%	.95%	.19%	.44%
12	.66%		.26%	
13	.86%		.39%	
14	1.24%		.54%	
15	1.59%		.80%	
16	2.20%	4.23%	1.41%	3.27%
17	2.96%		2.24%	
18	4.06%		3.09%	
19	5.31%		4.22%	
20	6.61%		5.37%	
21	8.84%	12.18%	7.58%	11.82%
22	10.33%		9.33%	
23	12.11%		11.76%	
24	13.95%		14.13%	
25	15.65%		16.28%	
26	17.96%	21.65%	18.36%	21.90%
27	20.00%		20.15%	
28	21.83%		22.14%	
29	23.42%		23.83%	
30	25.03%		25.00%	

Note. For Each Item = Percent of Missing Values in Each Item; Across Five Items = Percent of Missing Values Across a Grouping of five Items.

The results for the stratified samples were similar to the unstratified sample, such that the omission rates tended to increase throughout the test. The accession sources varied in their omission rates across the last five items on both Forms T1 and T2. The lowest and highest omission rates on T1 were observed for the ANG (17.36%) and USAFA (25.38%), respectively. On T2, the lowest omission rate was observed for the ANG (18.28%) and the highest omission rate was observed for the AECP (27.26%).

In general, the current pattern is consistent with the results reported by Morath et al. (2008) for Form S and by Carretta et al. (2016) for Form T for the unstratified sample. These reports did not include the omission and guessing rates for the stratified samples. Aguilar (2017) did not report omission rates or overt guessing.

3.9.6 Distractor Analysis

When examining the response frequencies, for six items on Form T1 and for eight items on T2 a distractor was chosen more frequently than a correct response. There were some items on Forms T1 and T2, for which a distractor was chosen less than 1%. Note that T1 Block 3 and T2 Block 4 are common and have a large number of problematic distractors. T1 Block 6 and T2 Block 6 are also common.

The pattern of results for the stratified samples was similar to the unstratified sample, such that all accession sources had at least one item for which a distractor was chosen more often than the correct response and several items with at least one infrequently chosen distractor.

In general, the current findings are consistent with the results reported by Carretta et al. (2016) for Form T. Aguilar (2017) did not report the distractor analysis. The pattern of results on BC can be explained by a disconnect between the instructions provided before the subtest and some of the items. For example, there were a few blocks touching on their backsides out of the test-takers' view, which was not illustrated in the instructions. Another reason for the presence of this many items with incorrectly selected distractors is the speeded nature of the subtest. When comparing the present results against the results for Form S reported by Morath et al. (2008) for the unstratified sample, no items on S1 and two items on S2 had a distractor chosen more often than a correct response.

3.9.7 Common Items

Fifteen items were common between Forms T1 and T2. They are displayed in Table 79. The table also displays the results of the independent samples *t*-tests comparing the *p*-values for these common items. As can be seen, 14 *t*-tests were statistically significant, with the effect sizes (Cohen's *d*) ranging from .03 to .46. The highest effect size of .46 was observed for T1-19 (Block 4) and T2-14 (Block 3). As mentioned earlier, there were a few issues with the BC subtest, concerning the instructions and the actual items. Also, the order of the blocks may render some items more susceptible to guessing and omission due to the subtest's speededness. Note that the *t*-tests were run for unstratified sample only. A difference in *p*-values was also examined for the stratified samples only. The differences between the *p*-values were small, ranging from $|.01|$ to $|.03|$ (see Attachment 1).

Table 79. BC Unstratified Sample Common Items Difficulty

T1 Block #	T1 Item #	T2 Block #	T2 Item #	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	Sig.	ES
2	T1-06	2	T2-06	.70	.46	.73	.44	.00	.08
	T1-07		T2-07	.74	.44	.60	.49	.00	.31
	T1-08		T2-08	.75	.43	.68	.47	.00	.14
	T1-09		T2-09	.84	.37	.84	.36	.42	.01
	T1-10		T2-10	.86	.35	.87	.33	.00	.03
3	T1-11	4	T2-16	.50	.50	.46	.50	.00	.08
	T1-12		T2-17	.19	.39	.18	.38	.00	.04
	T1-13		T2-18	.27	.44	.25	.43	.00	.05
	T1-14		T2-19	.37	.48	.33	.47	.00	.09
	T1-15		T2-20	.40	.49	.33	.47	.00	.14
4	T1-16	3	T2-11	.74	.44	.76	.43	.00	.04
	T1-17		T2-12	.57	.50	.59	.49	.00	.05
	T1-18		T2-13	.49	.50	.55	.50	.00	.14
	T1-19		T2-14	.73	.45	.90	.30	.00	.46
	T1-20		T2-15	.42	.49	.54	.50	.00	.24

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; Sig. = Significance Levels; ES = Effect Size.

The only technical report that included an examination of the common items was by Morath et al. (2008) for Form S. Similar to the current findings, the authors did not detect sizable differences among the common items on Form S.

3.9.8 Subgroup Differences

Table 80 summarizes the CTT effect sizes (Cohen's d) for the mean score difference among the gender, racial, and ethnic subgroups across Forms T1 and T2.

Regarding Female (F) and Male (M) subgroup comparisons, all items on T1 and all but one item on T2 (T2-05) had small effect sizes. On T1, the effect sizes ranged from .04 to .37, with M scoring higher than F on 30 items. On T2, the effect sizes ranged from .03 to .36, with M scoring higher than F on all 30 items.

Regarding Black (B) and White (W) subgroup comparisons, 17 items on T1 and 14 items on T2 showed moderate effect sizes. On T1, the effect sizes ranged from .17 to .64, with W scoring higher than B on 30 items. On T2, the effect sizes ranged from .17 to .64, with W scoring higher than B on all 30 items.

Regarding Asian (A) and White (W) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .00 to .18, with W scoring higher than A on 29 items. On T2, effect sizes ranged from .01 to .13, with W scoring higher than A on 28 items.

Regarding White Hispanic (WH) and White Non-Hispanic (WnH) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .03 to .20, with WnH scoring higher than WH on all 30 items. On T2, effect sizes ranged from .03 to .21, with WnH scoring higher than WH on all 30 items.

Regarding Black Hispanic (BH) and Black Non-Hispanic (BnH) subgroup comparisons, all effect sizes were small on both test forms. On T1, the effect sizes ranged from .00 to .15, with BnH scoring higher than BH on three items. On T2, the effect sizes ranged from .00 to .24, with BnH scoring higher than BH on one item.

Table 80. BC Unstratified Sample Mean Score Subgroup Differences - Cohen's *d*

Item #	T1														
	M_F	M_M	d	M_B	M_W	d	M_A	M_W	d	M_{WH}	M_{WnH}	d	M_{BH}	M_{BnH}	d
01	.39	.53	.28**	.34	.53	.37**	.50	.52	.05*	.53	.47	.13**	.35	.38	.06
02	.58	.72	.30**	.52	.71	.43**	.71	.71	.01 [†]	.72	.67	.10**	.53	.60	.15**
03	.42	.61	.39**	.35	.60	.50**	.59	.59	.01	.60	.54	.13**	.37	.37	.01
04	.62	.72	.21**	.55	.72	.36**	.68	.72	.08**	.72	.68	.08**	.56	.61	.09 [†]
05	.64	.77	.30**	.55	.78	.54**	.71	.78	.15**	.78	.73	.13**	.56	.58	.04
06	.62	.72	.22**	.55	.72	.38**	.72	.72	.00	.72	.71	.03	.57	.60	.07
07	.65	.78	.29**	.56	.78	.52**	.77	.78	.02	.78	.74	.10**	.57	.61	.08 [†]
08	.63	.79	.37**	.52	.79	.64**	.77	.79	.05*	.79	.74	.13**	.54	.59	.09 [†]
09	.80	.86	.17**	.73	.86	.35**	.85	.86	.04	.86	.83	.09**	.75	.74	.00
10	.79	.89	.30**	.72	.89	.51**	.86	.89	.07**	.89	.85	.11**	.73	.75	.04
11	.38	.55	.33**	.32	.54	.45**	.45	.54	.18**	.55	.47	.15**	.34	.34	.00
12	.12	.22	.25**	.10	.21	.29**	.20	.21	.04 [†]	.22	.17	.11**	.10	.11	.02
13	.19	.30	.24**	.15	.30	.34**	.26	.29	.08**	.30	.25	.11**	.16	.16	.01
14	.29	.40	.23**	.23	.40	.35**	.38	.40	.03	.40	.35	.10**	.24	.26	.06
15	.29	.44	.32**	.21	.44	.48**	.36	.44	.16**	.45	.38	.14**	.22	.25	.07
16	.69	.76	.16**	.57	.78	.48**	.76	.77	.04 [†]	.78	.73	.13**	.59	.62	.08 [†]
17	.48	.60	.26**	.36	.61	.51**	.56	.61	.11**	.62	.53	.17**	.37	.41	.09 [†]
18	.39	.52	.26**	.31	.52	.43**	.50	.52	.03	.52	.47	.10**	.32	.39	.15**
19	.70	.74	.10**	.56	.77	.47**	.69	.76	.17**	.77	.70	.17**	.57	.59	.03
20	.32	.46	.27**	.22	.47	.50**	.41	.46	.10**	.47	.39	.16**	.24	.28	.10*
21	.48	.57	.18**	.31	.60	.59**	.55	.59	.09**	.60	.51	.18**	.33	.33	.01
22	.56	.61	.10**	.38	.64	.55**	.59	.64	.11**	.65	.56	.20**	.40	.43	.06
23	.44	.50	.11**	.29	.52	.47**	.46	.52	.11**	.53	.43	.20**	.30	.35	.10*
24	.50	.53	.06**	.36	.56	.40**	.51	.55	.09**	.56	.48	.16**	.37	.39	.04
25	.31	.37	.13**	.21	.38	.36**	.34	.38	.08**	.39	.30	.18**	.22	.23	.01
26	.14	.20	.14**	.12	.20	.21**	.18	.20	.05*	.20	.16	.10**	.12	.11	.03
27	.05	.09	.13**	.04	.09	.17**	.08	.09	.04	.09	.07	.07**	.04	.05	.03
28	.30	.32	.04**	.19	.34	.33**	.30	.34	.07**	.34	.28	.14**	.20	.23	.08
29	.10	.16	.17**	.05	.16	.30**	.14	.16	.07**	.17	.12	.13**	.06	.09	.11*
30	.23	.28	.11**	.18	.29	.24**	.26	.29	.07**	.29	.26	.07**	.19	.19	.01

Item #	T2														
	M_F	M_M	d	M_B	M_W	d	M_A	M_W	d	M_{WH}	M_{WnH}	d	M_{BH}	M_{BnH}	d
1	.80	.87	.20**	.75	.87	.35**	.87	.87	.01	.87	.86	.05**	.76	.79	<u>.08[†]</u>
2	.61	.71	.23**	.52	.71	.42**	.70	.71	.02	.72	.67	.10**	.53	.60	<u>.14**</u>
3	.78	.82	.10**	.71	.83	.31**	.79	.83	.09**	.83	.81	.06**	.71	.72	<u>.02</u>
4	.88	.92	.13**	.85	.92	.27**	.88	.92	.15**	.92	.91	.06**	.86	.84	<u>.06</u>
5	.63	.80	.40**	.56	.79	.55**	.79	.78	<u>.00</u>	.79	.74	.12**	.57	.63	<u>.13**</u>
6	.66	.76	.22**	.60	.75	.35**	.78	.75	<u>.06**</u>	.75	.74	.03	.61	.66	<u>.09[†]</u>
7	.49	.64	.31**	.40	.64	.49**	.62	.63	.02	.64	.57	.14**	.40	.52	<u>.24**</u>
8	.56	.73	.36**	.44	.73	.64**	.70	.73	.05*	.73	.66	.15**	.45	.52	<u>.14**</u>
9	.80	.86	.16**	.75	.86	.30**	.84	.86	.06*	.86	.84	.07**	.76	.80	<u>.09[†]</u>
10	.81	.89	.25**	.74	.90	.50**	.89	.90	.04 [†]	.90	.87	.11**	.75	.79	<u>.08[†]</u>
11	.71	.78	.15**	.67	.78	.26**	.77	.78	.01	.78	.74	.08**	.68	.68	<u>.00</u>
12	.48	.63	.30**	.38	.63	.53**	.60	.63	.06*	.64	.57	.14**	.39	.46	<u>.15**</u>
13	.44	.59	.31**	.36	.59	.46**	.56	.59	.05*	.59	.55	.09**	.37	.42	<u>.10*</u>
14	.88	.91	.09**	.82	.92	.33**	.88	.92	.12**	.92	.88	.14**	.83	.83	<u>.01</u>
15	.43	.58	.31**	.33	.59	.52**	.54	.58	.09**	.59	.51	.15**	.34	.41	<u>.14**</u>
16	.35	.50	.30**	.29	.50	.43**	.44	.50	.13**	.51	.41	.21**	.31	.31	<u>.01</u>
17	.12	.20	.22**	.09	.20	.29**	.17	.20	.08**	.20	.17	.09**	.09	.14	<u>.17**</u>
18	.17	.27	.24**	.14	.27	.31**	.25	.27	.03	.27	.23	.10**	.14	.19	<u>.15**</u>
19	.26	.35	.20**	.18	.36	.38**	.33	.36	.06**	.36	.31	.11**	.19	.22	<u>.07</u>
20	.24	.37	.28**	.17	.37	.43**	.32	.37	.10**	.37	.31	.14**	.17	.22	<u>.12*</u>
21	.65	.69	.07**	.46	.72	.58**	.66	.72	.13**	.73	.66	.15**	.48	.54	<u>.13**</u>
22	.15	.18	.10**	.11	.19	.22**	.16	.19	.07**	.19	.16	.07**	.11	.13	<u>.07</u>
23	.25	.33	.16**	.24	.32	.17**	.31	.32	.02	.32	.30	.05*	.25	.27	<u>.06</u>
24	.19	.27	.19**	.09	.28	.45**	.23	.28	.11**	.29	.21	.17**	.10	.13	<u>.12*</u>
25	.50	.54	.09**	.37	.57	.41**	.52	.57	.10**	.57	.51	.13**	.37	.39	<u>.02</u>
26	.42	.45	.05**	.29	.47	.36**	.43	.47	.08**	.48	.40	.14**	.31	.32	<u>.03</u>
27	.23	.26	.08**	.14	.28	.31**	.24	.28	.08**	.28	.23	.11**	.15	.18	<u>.08[†]</u>
28	.24	.27	.05**	.11	.29	.41**	.25	.29	.08**	.29	.24	.13**	.12	.16	<u>.10*</u>
29	.20	.23	.09**	.10	.25	.34**	.23	.24	.04 [†]	.25	.19	.14**	.11	.14	<u>.10*</u>
30	.31	.33	.03**	.23	.35	.25**	.31	.34	.08**	.35	.31	.09**	.23	.26	<u>.07</u>

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

Note. M_F = Mean for Female Test-Takers; M_M = Mean for Male Test-Takers; M_B = Mean for Black/African American Test-Takers; M_W = Mean for White Test-Takers; M_A = Mean for Asian Test-Takers; d = Cohen's d . Underlined the effect sizes for the items for which the minority subgroups scored higher than their majority subgroups.

The general pattern of results was consistent among the stratified and unstratified samples, such that the majority of the items showed small effect sizes. The greatest number of items with moderate-to-large effect sizes was observed for B/W comparisons, with AFRES having the most items compared to the rest of the accession sources (22 on T1 and 20 on T2). AECP had two items with large effect sizes on T1 and one item on T2. Also, AECP had two items with large effect size on T2 for F/M. AFRES had one items with large effect size for BH/BnH on T2.

Table 81 displays the IRT DIF_{MH} results and Table 82 displays the DIF_{Lord} results for the same subgroups. There were three items flagged with moderate DIF on T1 but there was no convergence between the two DIF methods, so these results are likely spurious. On T2, however, T2-05 was flagged as moderate for F/W and T2-24 was flagged as moderate for B/W. No other convergence was observed.

Table 81. BC Unstratified Sample DIF_{MH}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.04	.97	.12	.01	.02	.07	.09	.32	.38	.06
02	.00	.80	.51	.26	.49	.00	.34	.30	.10	.05
03	.60	.48	.40	.13	.29	.50	.13	.32	.23	.15
04	.27	.81	.15	.17	.29	.00	.44	.88	.09	.66
05	.13	.03	.53	.10	.03	1.02	.33	.63	.19	.05
06	.12	.45	.32	.41	.08	.06	.27	.70	.42	.10
07	.08	.17	.30	.02	.18	.42	.19	.35	.06	.87
08	.43	.13	.20	.03	.02	.67	.26	.24	.11	.23
09	.10	.11	.05	.16	.28	.08	.04	.16	.02	.01
10	.38	.07	.11	.06	.07	.31	.44	.04	.04	.06
11	.58	.05	.65	.16	.14	.12	.02	.12	.10	.34
12	.56	.55	.05	.06	.11	.39	.08	.09	.01	.27
13	.27	.60	.17	.00	.21	.45	.19	.14	.24	.02
14	.16	.40	.13	.13	.27	.35	.80	.69	.62	.34
15	.41	.12	.48	.03	.02	.38	.16	.09	.09	.10
16	.28	.51	.06	.33	.08	.44	.21	.35	.45	.33
17	.05	.10	.24	.26	.07	.41	.06	.20	.04	.58
18	.05	.32	.26	.10	.52	.30	.42	.20	.09	.81
19	.39	.77	.62	.57	.37	.04	.16	.01	.02	.20
20	.10	.27	.14	.24	.24	.37	.15	.15	.08	.46
21	.41	.91	.06	.47	.37	.56	1.56	.37	.31	.11
22	.66	.89	.21	.59	.03	.42	.14	.19	.08	.24
23	.55	.77	.21	.61	.14	.01	.35	.14	.15	.31
24	.73	.51	.07	.46	.22	.03	1.38	.21	.29	.16
25	.15	.59	.04	.63	.26	.53	1.23	.22	.22	.11
26	.33	.95	.08	.04	.15	.56	1.15	.10	.21	.20
27	.18	1.04	.16	.08	.08	.57	.63	.06	.08	.03
28	.71	.73	.08	.49	.22	.71	1.99	.00	.10	.17
29	.15	.07	.01	.38	.53	.50	1.33	.25	.14	.16
30	.39	.19	.06	.03	.31	.57	.84	.13	.02	.04

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Table 82. BC Unstratified Sample DIF_{Lord}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.28	.41	.17	.06	.00	.24	.22	.35	.33	.03
02	.39	.16	.48	.20	.39	.14	.14	.31	.06	.01
03	.89	.26	.40	.08	.24	.51	.53	.21	.35	.27
04	.07	.52	.04	.32	.22	.11	.45	.78	.13	.79
05	.45	.44	.50	.03	.04	1.18	.54	.42	.11	.06
06	.03	.42	.42	.61	.12	.11	.47	.80	.49	.14
07	.44	.34	.30	.10	.12	.56	.27	.31	.04	.79
08	.90	.92	.12	.06	.00	.88	.98	.10	.19	.24
09	.15	.31	.18	.10	.33	.07	.53	.03	.14	.11
10	.77	.56	.12	.02	.01	.62	.62	.05	.12	.04
11	.58	.03	.59	.02	.16	.23	.91	.32	.16	.50
12	.57	.20	.17	.02	.13	.53	.51	.06	.11	.26
13	.27	.19	.07	.08	.21	.56	.15	.14	.14	.01
14	.06	.37	.27	.30	.23	.42	.15	.60	.42	.45
15	.59	.49	.46	.01	.07	.54	.49	.09	.17	.12
16	.34	.16	.18	.05	.06	.51	.06	.32	.44	.36
17	.14	.32	.18	.14	.06	.52	.01	.16	.06	.54
18	.15	.05	.28	.23	.51	.40	.17	.21	.08	.77
19	.69	.11	.57	.26	.38	.03	.08	.05	.06	.25
20	.27	.54	.14	.12	.26	.52	.46	.19	.18	.44
21	.33	.79	.09	.26	.26	.73	.67	.34	.21	.01
22	.73	.52	.20	.34	.05	.46	.70	.12	.19	.13
23	.70	.20	.18	.35	.18	.18	1.38	.32	.39	.14
24	1.00	.26	.04	.14	.17	.11	1.28	.32	.41	.20
25	.57	.24	.00	.32	.25	.68	.12	.15	.04	.26
26	.34	.96	.12	.10	.27	.89	.32	.06	.05	.35
27	.02	.71	.11	.05	.04	.68	.23	.09	.05	.14
28	1.12	.41	.02	.11	.25	.90	.74	.07	.11	.12
29	.01	.47	.08	.34	.56	.61	.32	.14	.20	.13
30	.67	.95	.04	.34	.37	1.00	.84	.05	.17	.26

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Aguilar (2017) and Carretta et al. (2016) did not include either CTT or IRT subgroup difference analyses in their report. Morath et al. (2008) included CTT Cohen's *d* results for Form S. The authors observed no items on S2 and two items on S1 with large effect sizes. However, there were items with moderate subgroup differences, with the highest number of items observed for B/W comparisons.

3.9.9 Conclusion

Overall, the items in the BC subtest are below average difficulty and of moderate discriminability. This subtest is speeded by design. There is some evidence to suggest that overt guessing occurred when examining the longstring option endorsement across the last five items. The removal of any one of the items would have a small effect on the internal consistency of the BC subtest. Internal

consistency is fairly high which may suggest redundancy. There were some distractors chosen more frequently than the correct response. As mentioned before, this could be due to an unfamiliar item presentation not used in the instructions and practice. Many distractors were chosen infrequently. Common items analysis showed small differences in item difficulty between Forms T1 and T2. Most items showed small-to-moderate (no large) Cohen's *d*, with the highest number of items with moderate Cohen's *d* observed for B/W comparisons. Similarly, most items showed small-to-moderate DIF across the subgroups.

3.10 Aviation Information (AI)

AI is the ninth cognitive subtest and is one of the five subtests that make up the Classification part of the AFOQT. It consists of 20 items with five multiple choice options. Test-takers are allowed 8 minutes to complete the test, for an average of 24 seconds per item.

3.10.1 Descriptives

Table 83 summarizes the means, standard deviations, skewness, and kurtosis for the unstratified sample across Forms T1 and T2. The item means ranged from .29 to .83 with an average of .47 on T1 and from .24 to .76 with an average of .45 on T2 and generally decrease throughout the test. Skewness tended to switch from negative to positive throughout the test, with the values remaining within the acceptable range.

Table 83. AI Unstratified Sample Descriptives

Item #	T1				T2			
	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
01	.59	.49	-.36	-1.87	.76	.43	-1.19	-.58
02	.74	.44	-1.08	-.83	.60	.49	-.41	-1.83
03	.83	.38	-1.72	.97	.69	.46	-.82	-1.32
04	.63	.48	-.55	-1.70	.59	.49	-.37	-1.87
05	.48	.50	.07	-1.99	.54	.50	-.15	-1.98
06	.50	.50	-.01	-2.00	.48	.50	.08	-1.99
07	.48	.50	.08	-1.99	.38	.49	.49	-1.76
08	.48	.50	.06	-2.00	.50	.50	.01	-2.00
09	.37	.48	.56	-1.69	.33	.47	.74	-1.46
10	.50	.50	.00	-2.00	.51	.50	-.03	-2.00
11	.52	.50	-.07	-2.00	.54	.50	-.15	-1.98
12	.55	.50	-.21	-1.96	.43	.50	.28	-1.92
13	.42	.49	.34	-1.89	.36	.48	.60	-1.64
14	.37	.48	.55	-1.70	.34	.47	.66	-1.56
15	.36	.48	.59	-1.65	.42	.49	.33	-1.89
16	.38	.48	.51	-1.74	.30	.46	.86	-1.26
17	.31	.46	.84	-1.30	.24	.43	1.22	-.52
18	.29	.45	.93	-1.14	.31	.46	.84	-1.29
19	.34	.47	.67	-1.54	.28	.45	1.00	-1.01
20	.30	.46	.89	-1.20	.33	.47	.74	-1.46

Note. *M* = Mean; *SD* = Standard Deviation. $N_{T1} = 36,039$ $N_{T2} = 32,904$.

The results for the stratified samples were similar to the unstratified sample in terms of decreasing means and skewness switching from negative to positive throughout the test. Some accession sources showed a higher number of positively skewed items (e.g., AECF, ROTC). The USAFA, ANG, and AFRES had the highest mean than any other accession source on T1 (.54, .54, and .53 respectively) and on T2 (.51, .52, and .49, respectively).

Carretta et al. (2016), Aguilar (2017), and Morath et al. (2008) do not report item-level central tendency statistics.

3.10.2 Difficulty

Table 84 summarizes the p -values and 2PL b parameters for the unstratified sample across Forms T1 and T2. The p -values ranged widely on both Forms T1 and T2, but in general, they tended to decrease throughout the test. None of the items showed a floor effect. One item on T1 showed a ceiling effect (T1-03 p -value = .83). The 2PL b parameters tended to switch from negative to positive throughout the test, with one item on T2 falling below the acceptable threshold (T2-01 b = -2.24).

Table 84. AI Unstratified Sample Difficulty

Item #	T1		T2	
	p -values	2PL b	p -values	2PL b
01	.59	-.48	.76	-2.24
02	.74	-1.86	.60	-.54
03	.83	-1.62	.69	-1.15
04	.63	-.98	.59	-.59
05	.48	.06	.54	-.16
06	.50	-.05	.48	.17
07	.48	.02	.38	.66
08	.48	.04	.50	.00
09	.37	.50	.33	1.22
10	.50	-.02	.51	-.06
11	.52	-.10	.54	-.17
12	.55	-.27	.43	.59
13	.42	.74	.36	.84
14	.37	.78	.34	1.30
15	.36	1.24	.42	.26
16	.38	.63	.30	.82
17	.31	.79	.24	1.28
18	.29	.81	.31	1.20
19	.34	.94	.28	1.08
20	.30	.82	.33	1.05

Note. p -value = Difficulty; 2PL b = IRT Difficulty Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample in terms of decreasing p -values and 2PL b parameters switching from negative to positive throughout the test. Form the CTT perspective, the USAFA, ANG, and AFRES scored higher on average than any other accession source, with T1 means averaging at .54, .54, and .53, respectively, and with T2

means averaging at .51, .52, and .49, respectively. No accession sources besides AECP had items with a floor effect. There were a few items across the accession sources that reached a ceiling effect (e.g., Item T1-03 for OTS-CIV). Notably, all items with a ceiling effect were towards the beginning of the test and all items with a floor effect were towards the end of the test across the accession sources. From the IRT perspective, three items on T1 (T1-01 to T1-03) and two items on T2 (T2-01 and T2-02) had 2PL b parameters below 2.00 for the USAFA. T1-02 had unacceptable 2PL b parameters for both the AECP and ROTC. T2-01 had unacceptable 2PL b parameters for OTS-AD, ROTC, and AFRES.

In general, the current findings are consistent with the results reported by Carretta et al. (2016) and Aguilar (2017) for Form T. When comparing the present pattern of findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T seems to be more difficult. The mean p -values for Form S1 and S2 were .52 and .50, respectively.

3.10.3 Discriminability

Table 85 summarizes the ITC's and 2PL a parameters for the unstratified sample across Forms T1 and T2. ITCs ranged from .31 to .60 with an average of .45 for T1 and from .29 to .57 with an average of .41 on T2. Nine items on T1 and 14 items on T2 had unacceptable 2PL a parameters.

Table 85. AI Unstratified Sample Discriminability

Item #	T1		T2	
	ITC	2PL a	ITC	2PL a
01	.44	.92	.29	.54
02	.31	.60	.43	.91
03	.37	1.21	.37	.79
04	.34	.60	.38	.69
05	.49	1.08	.54	1.62
06	.55	1.53	.32	.47
07	.60	1.94	.41	.82
08	.49	1.09	.41	.80
09	.55	1.41	.36	.64
10	.40	.78	.40	.80
11	.53	1.34	.52	1.38
12	.46	1.06	.32	.48
13	.31	.47	.40	.79
14	.41	.77	.33	.53
15	.32	.49	.57	1.68
16	.44	.92	.52	1.31
17	.53	1.32	.46	1.09
18	.55	1.49	.39	.76
19	.40	.78	.46	1.08
20	.54	1.39	.40	.77

Note. ITC = Item-Total Correlation; 2PL a = IRT Discriminability Parameter.

The pattern of results for the stratified samples was similar to the unstratified sample, such that there was a wide range of ITC's among the items and some items had low 2PL a parameters.

Interestingly, the number of items with high discriminability was the highest for OTS-CIV, ANG, and AFRES on T1 and for ANG and AFRES for T2.

In general, the current pattern of findings is consistent with the results reported by Carretta et al. (2016) and Aguilar (2017) for Form T. When comparing the present findings against Morath et al. (2008) and Aguilar (2017) for Form S, Form T was more discriminable ($T1 = .45$ and $T2 = .41$ vs. $S1 = .38$ and $S2 = .37$).

3.10.4 Internal Consistency if Item is Deleted

Table 86 summarizes Cronbach's alphas if an item is deleted for the unstratified sample across Forms T1 and T2. For reference, the AI subtest-level alphas for T1 and T2 were .79 and .75, respectively. The item-level alphas are high, ranging from .78 to .80 on T1 and from .72 to .74 on T2. All alphas were within the acceptable range. The removal of any one of the items would have a small effect on the internal consistency of the AI subtest.

Table 86. AI Unstratified Sample Internal Consistency if Item is Deleted

Item #	T1 α	T2 α
01	.79	.74
02	.79	.73
03	.79	.74
04	.79	.74
05	.78	.72
06	.78	.74
07	.78	.74
08	.78	.74
09	.78	.74
10	.79	.74
11	.78	.73
12	.79	.74
13	.80	.74
14	.79	.74
15	.79	.72
16	.79	.73
17	.78	.73
18	.78	.74
19	.79	.73
20	.78	.74

Note. α = Cronbach's Alpha.

The alphas across the accession sources varied. For Form T1, AECF was the only accession source with the subtest-level alphas of .68 – below the acceptable threshold. It did not look like removing any one item would improve the internal consistency for AECF. On Form T2, OTS-AD, AECF, and USAFA had unacceptable subtest-level alphas as well (.69, .66, and .68, respectively). There were not any one item the deletion of which would improve internal consistency. The rest of the accession sources had acceptable alphas.

Although the reports by Aguilar (2017), Carretta et al. (2016), and Morath et al. (2008) did not report the values for alpha if items were deleted, the Cronbach's alpha they reported for the subtests were similar to the alphas observed in this research.

3.10.5 Omission Rates and Overt Guessing

Table 87 summarizes the omission rates in a form of missing values for each item and for a grouping of five items for Forms T1 and T2. The longstring omission rate across the last five items was .33% for T1 and .39% for T2. In general, the omission rates seem to increase throughout the test. For T1, the average omission rate on the first five items was .15%, whereas the average on the last five items was .95%. For T2, these averages went from .21% to 1.16%. The longstring overt guessing for the last five items was .75% for T1 and .58% for T2.

Table 87. AI Unstratified Sample Omission Rates

Item #	T1		T2	
	For Each Item	Across Five Items	For Each Item	Across Five Items
01	.24%	.15%	.16%	.21%
02	.16%		.21%	
03	.05%		.16%	
04	.20%		.40%	
05	.10%		.14%	
06	.08%	.16%	.16%	.23%
07	.18%		.33%	
08	.14%		.26%	
09	.17%		.16%	
10	.21%		.25%	
11	.19%	.25%	.49%	.41%
12	.17%		.29%	
13	.27%		.31%	
14	.27%		.60%	
15	.34%		.39%	
16	.89%	.95%	.63%	1.16%
17	.65%		.78%	
18	.88%		1.12%	
19	1.06%		1.42%	
20	1.27%		1.85%	

Note. For Each Item = Percent of Missing Values in Each Item; Across Five Items = Percent of Missing Values Across a Grouping of five Items.

The results for the stratified samples were similar to the unstratified sample, such that the omission rates tended to increase throughout the test. The USAFA had the lowest omission rate in the grouping of the last five items (.19% for T1 and .31% for T2).

In general, the current pattern of findings is consistent with the results reported by Morath et al. (2008) for Form S and by Carretta et al. (2016) for Form T for the unstratified sample. These

reports did not include the omission and guessing rates for the stratified samples. Aguilar (2017) did not report omission rates or overt guessing.

3.10.6 Distractor Analysis

When examining the response frequencies, the correct response was chosen more frequently than a distractor for all items on both Forms T1 and T2, except for T1-17 and T1-18 and T2-09, T2-16, T2-17, and T2-19. T1-17 and T2-16 are common items. There were no items for which a distractor was chosen less than 1% of the time.

The pattern of results for the stratified samples was similar to the unstratified sample, such that most accession sources had at least one item for which a distractor was chosen more often than the correct response. There were two exceptions to this. ANG and AFRES on T1 endorsed all the keyed response options most frequently. On Form T2 however, ANG had one item for which the endorsement of a distractor was most frequent and AFRES had two items with a distractor being chosen most frequently. The accession sources with the most distractors chosen more often than the correct response was AECP. For several items across the accession sources, a distractor was chosen less than 1% (e.g., T1-02 for AECP; T2-05 for USAFA).

In general, the current findings are consistent with the results reported by Carretta et al. (2016) for Form T. Aguilar (2017) did not report the distractor analysis. When comparing the present results against the results for Form S reported by Morath et al. (2008) for the unstratified sample, all except for one common item on S1 and S2 had the correct answer chosen most frequently.

3.10.7 Common Items

Seven items were common between Forms T1 and T2. They are displayed in Table 88. The table also displays the results of the independent samples *t*-tests comparing the *p*-values for these common items. As can be seen, six *t*-tests were statistically significant, but the effect sizes (Cohen's *d*) were trivial, ranging from .02 to .07. Note that the *t*-tests were run for unstratified sample only. A difference in *p*-values was also examined for the stratified samples only. The differences between the *p*-values were small, ranging from |.00| to |.02| (see Attachment 1).

Table 88. AI Unstratified Sample Common Items Difficulty

T1 Item #	T2 Item #	M_{T1}	SD_{T1}	M_{T2}	SD_{T2}	Sig.	ES
T1-01	T2-02	.59	.49	.60	.49	.01	.02
T1-02	T2-01	.74	.44	.75	.43	.00	.04
T1-06	T2-05	.50	.50	.54	.50	.00	.07
T1-13	T2-12	.42	.49	.43	.50	.00	.03
T1-15	T2-14	.36	.48	.34	.47	.00	.03
T1-17	T2-16	.31	.46	.30	.46	.25	.01
T1-19	T2-20	.34	.47	.33	.47	.00	.03

Note. M_{T1} = Mean of T1; SD_{T1} = Standard Deviation of T1; M_{T2} = Mean of T2; SD_{T2} = Standard Deviation of T2; Sig. = Significance Levels; ES = Effect Size.

The only technical report that included an examination of the common items was by Morath et al. (2008) for Form S. Similar to the current findings, the authors did not detect sizable differences among the common items on Form S.

3.10.8 Subgroup Differences

Table 89 summarizes the CTT effect sizes (Cohen's *d*) for the mean score difference among the gender, racial, and ethnic subgroups across Forms T1 and T2.

Regarding Female (F) and Male (M) subgroup comparisons, seven items on T1 and eight on T2 showed moderate effect sizes. On T1, the effect sizes ranged from .04 to .53, with M scoring higher than F on all 20 items. On T2, the effect sizes ranged from .03 to .51, with M scoring higher than F on all 20 items.

Regarding Black (B) and White (W) subgroup comparisons, seven items on T1 and six on T2 showed moderate effect sizes. On T1, the effect sizes ranged from .17 to .64, with W scoring higher than B on all 20 items. On T2, the effect sizes ranged from .17 to .64, with W scoring higher than B on all 20 items.

Regarding Asian (A) and White (W) subgroup comparisons, all items on T1 and 19 items on T2 showed small effect sizes. T2-06 showed moderate effect size for A/W. On T1, the effect sizes ranged from .00 to .33, with W scoring higher than A on all 20 items. On T2, the effect sizes ranged from .01 to .42, with W scoring higher than A on all 20 items.

Regarding White Hispanic (WH) and White Non-Hispanic (WnH) subgroup comparisons, all items showed small effect sizes. On T1, the effect sizes ranged from .03 to .24, with WnH scoring higher than WH on all 20 items. On T2, the effect sizes ranged from .03 to .32, with WnH scoring higher than WH on all 20 items.

Regarding Black Hispanic (BH) and Black Non-Hispanic (BnH) subgroup comparisons, all effect sizes were small. On T1, the effect sizes ranged from .00 to .15, with BnH scoring higher than BH on eight items. On T2, effect sizes ranged from .00 to .24, with BnH scoring higher than BH on five items.

Table 89. AI Unstratified Sample Mean Score Subgroup Differences - Cohen's *d*

Item #	T1														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
01	.44	.64	.43**	.45	.63	.38**	.53	.63	.21**	.52	.64	.24**	.44	.46	.05
02	.64	.77	.32**	.65	.76	.25**	.71	.76	.10**	.73	.76	.08**	.68	.65	<u>.05</u>
03	.73	.86	.33**	.68	.86	.49**	.81	.86	.13**	.81	.86	.16**	.72	.69	<u>.07</u>
04	.56	.66	.21**	.48	.67	.40**	.58	.67	.18**	.61	.67	.14**	.48	.50	.04
05	.30	.55	.51**	.27	.53	.53**	.40	.53	.25**	.45	.54	.18**	.32	.29	<u>.07</u>
06	.35	.56	.43**	.34	.54	.40**	.42	.54	.25**	.46	.55	.17**	.34	.36	.04
07	.29	.55	.53**	.30	.52	.45**	.40	.52	.24**	.42	.53	.22**	.34	.31	<u>.05</u>
08	.35	.53	.37**	.34	.53	.39**	.39	.53	.27**	.44	.53	.19**	.36	.35	<u>.02</u>
09	.25	.41	.32**	.24	.40	.34**	.28	.40	.25**	.32	.41	.19**	.25	.24	<u>.00</u>
10	.45	.52	.13**	.44	.52	.17**	.48	.52	.08**	.48	.52	.09**	.44	.44	<u>.00</u>
11	.34	.58	.50**	.35	.56	.44**	.44	.56	.23**	.48	.57	.18**	.37	.36	<u>.03</u>
12	.44	.59	.30**	.38	.60	.44**	.47	.59	.26**	.51	.60	.20**	.40	.40	<u>.02</u>
13	.26	.48	.45**	.32	.45	.26**	.34	.45	.21**	.40	.45	.10**	.34	.33	<u>.03</u>
14	.29	.40	.22**	.22	.41	.39**	.29	.41	.25**	.32	.41	.18**	.23	.23	.00
15	.28	.38	.21**	.21	.40	.40**	.24	.40	.33**	.31	.41	.21**	.21	.22	.02
16	.25	.42	.34**	.24	.41	.34**	.36	.40	.09**	.33	.41	.16**	.28	.25	<u>.08</u> [†]
17	.16	.36	.45**	.16	.35	.39**	.23	.34	.24**	.25	.35	.22**	.16	.17	.05
18	.19	.32	.30**	.18	.32	.31**	.23	.32	.19**	.26	.32	.14**	.18	.18	.00
19	.30	.36	.12**	.25	.37	.25**	.27	.37	.21**	.31	.37	.13**	.27	.25	<u>.04</u>
20	.17	.34	.39**	.16	.34	.40**	.18	.34	.33**	.24	.35	.23**	.16	.17	.02
Item #	T2														
	<i>M_F</i>	<i>M_M</i>	<i>d</i>	<i>M_B</i>	<i>M_W</i>	<i>d</i>	<i>M_A</i>	<i>M_W</i>	<i>d</i>	<i>M_{WH}</i>	<i>M_{WnH}</i>	<i>d</i>	<i>M_{BH}</i>	<i>M_{BnH}</i>	<i>d</i>
1	.65	.79	.35**	.68	.78	.23**	.73	.78	.10**	.75	.78	.06**	.71	.68	<u>.07</u>
2	.45	.65	.43**	.46	.64	.37**	.52	.64	.25**	.55	.65	.21**	.44	.47	.06
3	.54	.74	.43**	.50	.74	.53**	.63	.73	.23**	.67	.74	.16**	.56	.52	<u>.08</u>
4	.49	.63	.27**	.49	.62	.25**	.51	.62	.23**	.55	.62	.14**	.53	.50	<u>.07</u>
5	.35	.60	.51**	.33	.59	.52**	.44	.59	.30**	.51	.59	.17**	.36	.36	<u>.00</u>
6	.39	.51	.24**	.33	.53	.41**	.32	.53	.42**	.39	.55	.32**	.32	.34	.04
7	.27	.42	.32**	.26	.41	.31**	.36	.41	.11**	.34	.42	.15**	.29	.27	<u>.06</u>
8	.40	.53	.28**	.36	.53	.35**	.46	.53	.15**	.47	.53	.12**	.37	.37	<u>.02</u>
9	.26	.35	.18**	.26	.35	.18**	.31	.34	.08**	.32	.35	.06**	.25	.26	.03
10	.46	.52	.13**	.44	.53	.17**	.47	.53	.11**	.48	.53	.10**	.49	.44	<u>.09</u> *
11	.37	.60	.46**	.31	.60	.58**	.42	.60	.35**	.48	.61	.26**	.37	.33	<u>.08</u> [†]
12	.27	.49	.44**	.32	.46	.27**	.34	.46	.24**	.40	.47	.13**	.36	.33	<u>.06</u>
13	.27	.39	.23**	.23	.39	.34**	.28	.39	.22**	.33	.40	.15**	.25	.24	<u>.02</u>
14	.28	.37	.19**	.20	.38	.38**	.21	.39	.36**	.28	.40	.23**	.21	.21	.02
15	.25	.48	.46**	.25	.46	.43**	.38	.46	.16**	.39	.46	.14**	.27	.26	<u>.01</u>
16	.15	.36	.45**	.15	.34	.41**	.22	.34	.26**	.25	.35	.20**	.13	.17	.10*
17	.11	.28	.41**	.14	.27	.29**	.19	.27	.17**	.22	.27	.11**	.15	.15	<u>.00</u>
18	.23	.33	.23**	.23	.33	.20**	.25	.33	.16**	.29	.33	.08**	.26	.24	<u>.05</u>
19	.17	.32	.34**	.18	.30	.26**	.22	.30	.17**	.25	.31	.12**	.19	.19	<u>.01</u>
20	.28	.34	.13**	.22	.36	.28**	.26	.35	.19**	.30	.36	.13**	.26	.24	<u>.06</u>

[†]*p* ≤ .100; **p* ≤ .050; ***p* ≤ .010; unmarked are non-significant *p*.

Note. *M_F* = Mean for Female Test-Takers; *M_M* = Mean for Male Test-Takers; *M_B* = Mean for Black/African American Test-Takers; *M_W* = Mean for White Test-Takers; *M_A* = Mean for Asian Test-Takers; *d* = Cohen's *d*. Underlined the effect sizes for the items for which the minority subgroups scored higher than their majority subgroups.

The general pattern of results was consistent among the stratified and unstratified samples, such that the majority of the items for the ethnic subgroups showed small effect sizes, with a few exceptions that slightly crossed the moderate threshold (e.g., Item 15 on T1 had a Cohen's d of .44 for the BH/BnH subgroups for ANG; Item 1 had a Cohen's d of .49 for the WnH/WH subgroups for AFRES). Item 5 on T1 was one of the two items to show a large effect size (A/W Cohen's d = .82 for AECP; B/W Cohen's d = .86 for AFRES). T1-01 showed a large effect size (A/W Cohen's d = .82 for AECP). T2-11 showed a large effect size for B/W subgroups for AECP and for AFRES (Cohen's d = .89 and Cohen's d = .80, respectively). There were three items that showed large effect sizes for AECP (T2-03 and T2-05 for B/W subgroups and T2-04 for the A/W subgroups), however, as mentioned before, the AECP sample sizes were rather small and therefore the interpretation of these analyses is to be done with caution.

Table 90 displays the IRT DIF_{MH} results and Table 91 displays the DIF_{Lord} results for the same subgroups. T1-10 and T1-19 showed moderate effect sizes for F/M subgroup comparisons across the two DIF approaches. Similarly, T2-06 and T2-14 showed moderate effect sizes for A/W subgroup comparisons across the two DIF approaches. Finally, T2-10 and T2-20 showed moderate effect sizes for F/M subgroup comparisons.

Table 90. AI Unstratified Sample DIF_{MH}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.27	.10	.19	.35	.33	.27	.32	.49	.17	.28
02	.07	.08	.39	.21	.25	.45	.02	.02	.36	.39
03	.04	.87	.43	.10	.22	.55	.91	.03	.23	.04
04	.67	.56	.06	.08	.22	.30	.36	.01	.11	.04
05	.65	.93	.01	.06	.28	.84	.52	.14	.02	.16
06	.04	.15	.21	.20	.18	.50	.61	1.22	.98	.33
07	.53	.11	.44	.02	.16	.24	.19	.65	.06	.22
08	.11	.03	.12	.08	.11	.38	.07	.42	.06	.03
09	.61	.34	.07	.01	.13	.86	.67	.69	.28	.08
10	1.17	.77	.68	.36	.05	1.11	.89	.57	.13	.31
11	.49	.10	.26	.07	.09	.62	.94	.40	.53	.36
12	.53	.37	.13	.15	.13	.57	.11	.28	.08	.29
13	.86	.11	.16	.10	.05	.60	.21	.09	.13	.08
14	.73	.62	.29	.29	.11	.77	.76	1.15	.72	.29
15	.54	.97	.92	.47	.08	.64	.24	.87	.18	.17
16	.16	.03	.78	.00	.37	.76	.45	.11	.31	.78
17	.58	.13	.11	.28	.44	.72	.12	.32	.14	.06
18	.66	.58	.55	.38	.13	.65	.57	.22	.30	.04
19	1.24	.11	.13	.10	.16	.24	.41	.34	.08	.11
20	.17	.31	.70	.30	.13	1.18	.09	.00	.06	.06

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Table 91. AI Unstratified Sample DIF_{Lord}

Item #	T1					T2				
	F/M	B/W	A/W	WH/WnH	BH/BnH	F/M	B/W	A/W	WH/WnH	BH/BnH
01	.38	.08	.09	.32	.31	.24	.56	.58	.41	.32
02	.05	.64	.61	.42	.26	.46	.06	.13	.28	.35
03	.26	.74	.30	.13	.23	.56	.85	.08	.08	.09
04	.78	.02	.24	.16	.23	.32	.57	.02	.04	.07
05	.85	.86	.16	.02	.26	.87	.89	.42	.04	.10
06	.38	.07	.13	.03	.13	.56	.30	1.15	.79	.33
07	.99	.38	.07	.22	.17	.05	.19	.64	.07	.21
08	.08	.02	.27	.08	.09	.35	.03	.42	.20	.00
09	.17	.14	.20	.08	.13	.85	.92	.82	.49	.11
10	1.26	1.27	.81	.52	.03	1.11	1.02	.62	.25	.32
11	.78	.24	.04	.04	.06	.63	1.21	.68	.53	.36
12	.33	.22	.18	.10	.14	.61	.44	.12	.20	.31
13	.61	.73	.09	.39	.07	.55	.08	.08	.03	.09
14	.77	.22	.18	.14	.12	.79	.41	1.03	.51	.26
15	.82	.29	.71	.20	.06	.76	.54	.35	.05	.09
16	.01	.13	.77	.06	.39	.97	.75	.38	.33	.73
17	.91	.40	.21	.36	.47	.93	.01	.14	.17	.01
18	.17	.20	.15	.17	.09	.56	.77	.26	.44	.05
19	1.34	.70	.03	.28	.11	.26	.32	.17	.09	.11
20	.44	.47	.93	.36	.16	1.14	.29	.10	.11	.03

Note. F/M = Effect Sizes for Female and Male Test-Takers; B/W = Effect Sizes for Black/African American and White Test-Takers; A/W = Effect Sizes for Asian and White Test-Takers; WH/WnH = Effect Sizes for White Hispanic and White Non-Hispanic Test-Takers; BH/BnH = Effect Sizes for Black Hispanic and Black Non-Hispanic Test-Takers.

Aguilar (2017) and Carretta et al. (2016) did not include either CTT or IRT subgroup comparison analyses in their reports. Morath et al. (2008) included CTT Cohen's *d* results for Form S. The authors observed large effect sizes for two items on S1 and for one item on S2, both for the B/W subgroup comparison. The examination of the Cohen's *d* tables in the appendices for the Morath et al. report revealed that nearly half of the items across almost all the subgroups had either moderate or large effect sizes.

3.10.9 Conclusion

Overall, the items in the AI subtest seem to be of average difficulty and of low-to-moderate discriminability. Item omission rates and longstring option endorsement across the last five items do not suggest speededness or overt guessing. The removal of any one of the items would have a small effect on the internal consistency of the subtest. There were a few distractors chosen more frequently than the correct response. No distractors were chosen infrequently. Common items analysis showed small differences in item difficulty between Forms T1 and T2. Most items showed small-to-moderate (no large) Cohen's *d* and DIF across the subgroups.

4.0 CONCLUSION

Based on this pattern of results for the AFOQT Form T item-level analyses, we recommend the following. First, most of the items are of below average difficulty. The next iteration of the AFOQT should explicitly specify what level of difficulty is preferred and develop new items or revise the existing items accordingly. Second, most of the items and their subtests are of low or moderate discriminability. The items should be revised to differentiate between, say, low ability level and lower ability level (or, between high ability level and higher ability level). It is usually fairly easy for a test item to differentiate between high and low ability tests takers, however it is harder to design items that would differentiate between low and lower end of the ability spectrum. Third, many items between the parallel forms and across the subtests seem to have problematic distractors, such that the distractors are either chosen more frequently than the keyed response or are chosen infrequently. The problematic distractors should be revised. Fourth, based on the inflated Cronbach's alphas, some subtests may be shortened without sacrificing measurement precision (e.g., IC, TR). Finally, there were items with large Cohen's d and DIF effect sizes across the gender, racial, and ethnic subgroups. The items should be reviewed qualitatively to ensure that there are no obvious culturally-laden items. Any problematic items should be replaced (and subsequently reviewed for gender, racial, and ethnic sensitivity) to potentially reduce mean score subgroup differences.

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 items needs to be based on the holistic interpretation of the data resulting from: Subtest-level analyses (Walsh et al., 2022), composite-level analyses, criterion-related validities, adverse impact rates and subgroup qualification rates, critical officer job analyses (Kantrowitz et al., in press), and analyses for non-cognitive subtests, including the SJT (Walsh et al., 2021) and the Self-Description Inventory (Woolley et al., under review).

5.0 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 Assessment Branch.
- Allison, P. D. (2001). *Missing data*. Sage publications.
- Carretta, T. R., Rose, M. R., & Trent, J. D. (2016). *Air Force Officer Qualifying Test Form T: initial item-, test-, factor-, and composite-level analyses*, AFRL-RH-WP-TR-2016-0093. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Systems Directorate, Supervisory Control and Cognition Branch.
- Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology*, 84(4), 610.
- Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological Bulletin*, 112(1), 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), 292-307.
- 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.
- Fedrigo, 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.
- Gao, X. (2019). *A comparison of six DIF detection methods*. University of Connecticut.: Master's Thesis.
- 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 composite*, AFRL-RH-WP-TR-2022-xxxx. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
- Kim, J. (2010). *Controlling Type I error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjustment of multiple item testing*. Georgia State University.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling*. NY: Guilford.
- Kline, T. J. (2005). Classical test theory: Assumptions, equations, limitations, and item analyses. *Psychological testing: A practical approach to design and evaluation* (pp. 91-105). Thousand Oaks, CA: Sage.
- Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Masters, G. N., & Wright, B. D. (1997). The partial credit model. In *Handbook of modern item response theory* (pp. 101-121). NY: Springer.
- 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.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* 3rd ed. NY: McGraw-Hill.
- Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, 81(2), 93-103.
- Reise, S. P., & Waller, N. G. (2002). Item response theory for dichotomous assessment data. 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.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147-177.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350-353.

- Shore, C. W., Haight, N., & Martinez, L. (2020). *Identify potential I/O or non I/O psychology assessment tools/Methods: AFOQT methods to reduce adverse impact*, unpublished technical report. Randolph AFB, TX Air Force Personnel Center, Strategic Research and Assessment Branch.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important?. *Journal of Applied Psychology*, 89(3), 497-508.
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5), pp. 481-498). Boston, MA: Pearson.
- Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, 26(2), 161-176.
- US Equal Employment Opportunity Commission, Civil Service Commission. (1978). *Uniform Guidelines on Employee Selection*, Washington, DC.
- Walsh, J. L., Woolley, M. R., Brady, M. F., & Carretta, T. R. (2022). Air Force Officer Qualifying Test (AFOQT) Form T evaluation: Subtest-level analyses, AFRL-RH-WP-TR-2022-0022. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
- Walsh, J. L., Woolley, M. R., Brady, M. F., Melick, S. R., & Carretta, T. R. (2021). Air Force Officer Qualifying Test (AFOQT) Form T evaluation: The Situational Judgment Test, AFRL-RH-WP-TR-2022-0080. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
- Woolley, M. R., Walsh, J. L., Mann, K. J., Wilson, R. T., & Carretta, T. R. (in press). Self-Description Inventory-Officer (SDI-O): Item-, facet-, and domain-level analyses, AFRL-RH-WP-TR-2022-xxxx. Wright-Patterson AFB, OH: s711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.
- Zickar, M. J. (2012). A review of recent advances in item response theory. In J. J. Martocchio, A. Joshi, & H. Liao (Eds.), *Research in Personnel and Human Resources Management. Vol. 31* (pp. 145-176). Bingley: Emerald Group Publishing.
- Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*, 37-59.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

$\geq$	Greater than or equal to
$\leq$	Less than or equal to
1PL	One-parameter logistic IRT model
2PL	Two-parameter logistic IRT model
3PL	Three-parameter logistic IRT model
ABM	Air Battle Manager
AD-Other	Active Duty Other Accession Source
AECF	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
AI	Aviation Information
AIAN	American Indian/Alaska Native
ANG	Air National Guard
AR	Arithmetic Reasoning
ASVAB	Armed Services Vocational Aptitude Battery
B	Black
BC	Block Counting
BH	Black Hispanic
B/W	Black/White
BnH	Black Non-Hispanic
CIV-Other	Civilian Other Accession Source
Cohen's d (d)	Effect Size
CSO	Combat Systems Officer
CTT	Classical Test Theory
DIF	Differential Item Functioning
DIF _{MH}	Mantel-Haenszel (DIF Method = DIF _{MH})
DoD	Department of Defense
F	Female
GPCM	Generalized Partial Credit Model
GRM	Graded Response Model
GS	General Science
HF	Hidden figures
IC	Instrument Comprehension
IER	Insufficient Effort Responding
IRT	Item Response Theory
IST	Infoscitex
ITC	Item-Total Correlation
M	Mean
M	Male
MAR	Missing at Random

MCAR	Missing Completely at Random
MK	Math Knowledge
MNAR	Missing Not at Random
<i>N</i>	Sample Size
NHPI	Native Hawaiian/Other Pacific Islander
OTS	Officer Training School
OTS-AD	Officer Training School-Active Duty
OTS-CIV	Officer Training School-Civilian
<i>p</i>	Probability value
PCA	Principal Components Analysis
PII	Personally Identifiable Information
PS	Physical Science
<i>p</i> -value	Proportion of correctly endorsed items in the test
RB	Rotated Blocks
RC	Reading Comprehension
ROTC	Reserve Officer Training Corps
RPA	Remotely-Piloted Aircraft
<i>SD</i>	Standard Deviation
SDI-O	Self-Description Inventory - Officers
<i>Sig.</i>	Significance
SJT	Situational Judgment Test
<i>t</i>	A symbol used to denote the <i>t</i> -statistic
T1	Air Force Officer Qualifying Test Form T Form 1
T2	Air Force Officer Qualifying Test Form T Form 2
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
W	White
WH	White Hispanic
WK	Word Knowledge
WnH	White Non-Hispanic