AFRL-RH-WP-TR-2021-0080

AIR FORCE OFFICER QUALIFYING TEST (AFOQT) FORM T: PSYCHOMETRIC EVALUATION OF THE SITUATIONAL JUDGMENT TEST

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

Thomas R. Carretta
711 HPW/RHBC

December 2021
Interim Report

DISTRIBUTION 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

AFRL-RH-WP-TR-2021-0080 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//	//signature//
THOMAS R. CARRETTA, PhD Work Unit Manager Performance Optimization Branch Airman Biosciences Division	R. ANDY MCKINLEY, DR-III, PhD Core Research Area Lead Cognitive and Physical Performance Performance Optimization Branch Airman Biosciences Division

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE ***(DD-MM-YY)*** 25-11-21	**2. REPORT TYPE** Interim	**3. DATES COVERED** ***(From - To)*** 1 April 2021 – 24 June 2021

4. TITLE AND SUBTITLE Air Force Officer Qualifying Test (AFOQT) Form T: Psychometric Evaluation of the Situational Judgment Test	**5a. CONTRACT NUMBER** FA8650-20-F-4094
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Julia L. Walsh[a], Montana R. Woolley[a], Michael F. Brady[a], Sarah R. Melick[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) Infoscitex, a DCS Company[a] 4027 Colonel Glenn Highway, Suite 210 Dayton, OH 45431-1672	**8. PERFORMING ORGANIZATION REPORT NUMBER**

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		**10. SPONSORING/MONITORING AGENCY ACRONYM(S)** 711 HPW/RHBC
Air Force Materiel Command[b] Air Force Research Laboratory 711th Human Performance Wing Airman Systems Directorate Airman Biosciences Division Performance Optimization Branch Wright-Patterson AFB, OH 45433	Infoscitex Corporation 4027 Colonel Glenn Highway Suite 210 Dayton, OH 45431-1672	**11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)** AFRL-RH-WP-TR-2021-0080

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A: Approved for public release.

13. SUPPLEMENTARY NOTES

Report contains color. AFRL-2022-0175, cleared 13 January 2022

14. ABSTRACT

The Air Force Officer Qualifying Test (AFOQT) is used to qualify applicants for officer commissioning or for aircrew training as pilots, combat system operators, air battle managers, and remotely-piloted aircraft pilots. This technical report summarizes an item-level and subtest-level psychometric evaluation of the Situational Judgment Test (SJT) included in the AFOQT Form T. Overall, the SJT was found to be relatively easy, with low-to-moderate ability to differentiate among the test-takers with varying standings on the latent trait, and low-to-moderate internal consistency. Analyses at the item-level revealed mostly small mean score subgroup differences, except for one response on Form T1 and two responses on Form T2. Analyses at the subtest-level revealed small to moderate mean score subgroup differences, with Black/African-American test-takers having the largest effect sizes, followed by Asian test-takers. Test-retest reliability and alternate forms reliability were acceptable. The differences between the parallel Forms T1 and T2 were practically nonsignificant, rendering equating unnecessary. The SJT was found to correlate stronger with the AFOQT cognitive subtests than with the personality subtest. The SJT was roughly at a 9-10 reading grade level. There was no evidence of test security breach. Recommendations for improvement are provided.

15. KEY WORDS

Situational Judgment Test, SJT, AFOQT, Air Force Officer Qualifying Test, officership assessment

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:** SAR	**18. NUMBER OF PAGES** 116	**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

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

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

Overall, the SJT was found to be relatively easy, with low-to-moderate ability to differentiate among the test-takers with varying standings on the latent trait, and low-to-moderate internal consistency. Analyses at the item-level revealed mostly small mean score subgroup differences, except for one response on T1 and two responses on T2. Analyses at the subtest-level revealed small to moderate mean score subgroup differences, with Black test-takers having the largest effect sizes, followed by Asian test-takers. Test-retest reliability and alternate forms reliability were acceptable. The differences between the parallel forms T1 and T2 were practically nonsignificant, rendering equating unnecessary. The SJT was found to correlate stronger with the AFOQT cognitive subtests than with the personality subtest. The SJT was roughly at a 9-10 reading grade level. There was no evidence of test security breach.

Although some of these findings were consistent with the literature, especially for the current SJT format, more recent research provides guidelines for how to improve the psychometric properties of the SJT in the future iterations. First, it is recommended that a shorter list of critical officership competencies be identified to guide test content. Previous AFOQT SJT research identified seven competencies, of which six were used to generate situations based on the Critical Incidents Technique. These situations were not balanced across the six competencies. This resulted in the SJT being multidimensional, which likely caused downward bias in some statistics and rendered others uninterpretable. To resolve this issue, a construct-driven approach focusing on a small number of competencies is recommended.

Second, given that SJT's are methods and not constructs, format changes may improve not only their psychometric properties, but also may further reduce mean score subgroup differences. Therefore, it is recommended that the knowledge-based instructions ("what should you do?") be replaced with the behavioral tendency instructions ("what would you do?"). Additionally, it is recommended to eliminate job-specific (militarized) item content in favor of job-generic (de-militarized) item content.

Finally, it is recommended that a computer-version of the SJT be considered. This would allow for improved data collection, and the collection of meta-data such as response latency, which is useful in reducing data noise and detecting Insufficient Effort Responding. As the AFOQT evolves, consideration may be given to multimedia presentation of the SJT to reduce the verbal loading of the SJT.

While these results and recommendations are meant to inform decisions regarding future iterations of the SJT, we caution that the ultimate decisions should be based on the holistic results obtained from the item-level and composite-level analysis of the AFOQT cognitive subtests (Kantrowitz et al., under review; Walsh et al., under review, a; b) and the other non-cognitive subtest, Self-Description Inventory-Officer (SDI-O) (Woolley et al., in progress).

1.0 INTRODUCTION

The SJT is a non-cognitive subtest that was introduced to the AFOQT Form T in 2014 with three objectives: (1) to expand the competencies measured by the AFOQT, (2) to incrementally improve criterion-related validity above and beyond the AFOQT cognitive subtests, and (3) to meet the United States Air Force (USAF) diversity and inclusion (D&I) goals by reducing mean score differences for historically underrepresented demographic subgroups (Barron, 2013; Lentz et al., 2009a; 2009b). The current technical report summarizes the psychometric evaluation of the SJT administered between 2014 and 2020. The review and resulting recommendations aim to improve the SJT's performance across all three objectives. The following sections briefly describe the history of the AFOQT and the initial and current SJT validation efforts.

1.1 Brief History of the AFOQT

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

1.1.1. AFOQT Form T

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

Table 1. AFOQT Form T Subtests

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

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

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

USAF cadets, enlisted personnel, and civilians from a variety of accession sources including Officer Training School-Civilian (OTS-CIV), Officer Training School-Active Duty (OTS-AD), AECP, United State Air Force Academy (USAFA), 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 Combat Systems Officer (CSO), Air Battle Manager (ABM), fighter or mobility pilot, and Remotely Piloted aircraft (RPA) pilot. Each applicant currently has an opportunity to take the assessment up to three times (i.e., the first attempt and two re-takes) with an opportunity to take the assessment a fourth 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).

1.1.2. History of the SJT and Initial Validation Effort

Following a literature review and workshops with over 4,000 USAF O1-O3 officers (Lentz et al., 2009a; 2009b), it was determined that the AFOQT required expanded competency coverage. The SJT was recommended as one of the viable options to meet that goal.

Realizing that SJT's are methods and not constructs (Schmitt & Chan, 2006), the initial SJT development aimed to measure seven core constructs (i.e., competencies): (1) Displaying Integrity, (2) Ethical Behavior and Professionalism, (3) Leading Others, Decision-Making and Managing Resources, (4) Communication Skills, (5) Leading Innovation, (6) Mentoring Others, and (7) Pursuing Personal and Professional Development. Eventually, Pursuing Personal and Professional Development was dropped and the final SJT focused on the other six competencies. For competency definitions please refer to Appendix A.

The SJT development performed by Barron (2013) followed a multiphasic approach. Phase I involved generating situations that officers would likely face in their careers. Barron followed the Critical Incidents Technique (CIT) using focus groups with 79 O-3s from Randolph Air Force Base (AFB) and Lackland AFB. Phase II involved focus groups with 22 O-3s (i.e., Air Force Captains) and 31 O-2s (i.e., Air Force First Lieutenants) and online surveys with 100 O-1s (i.e., Air Force Second Lieutenants) at Lackland AFB to elicit behavioral responses to the situations. Phase III involved generating an SJT scoring key for the situations by surveying 264 high-performing O-3s from Squadron Officer School. Eighty-six (86) SJT items, each with five to seven response options that aimed to measure the six aforementioned competencies, made up the original item bank. The initial validation effort performed with enlisted USAF personnel undergoing Basic Military Training (BMT) affirmed the potential benefits of including the SJT on the AFOQT. Specifically, it was demonstrated that the SJT (1) had appropriate psychometric properties; (2) was viewed by the USAF members as face valid and fair; and (3) showed a reduction in adverse impact when compared to the traditional cognitive tests (Barron, 2013). Based on that research, the items with the strongest psychometric properties with five response options were selected for inclusion in the AFOQT Form T. The final version of the SJT consisted of 25 items on T1 and 25 items on T2. Each parallel form contained 12 unique items and 13 common items.

1.1.3. Current SJT Administration

When taking the AFOQT, test-takers are presented with 25 SJT situations with five response options each. Test-takers are asked to select both the most effective (ME) and the least effective (LE) action for each situation. Thirty-five (35) minutes are allotted to complete this subtest. Note that the SJT is not part of any operational composites. This information is included in the AFOQT marketing materials (i.e., test pamphlets).

1.2 Present Research

The Air Force Personnel Center Strategic Research and Assessments Branch (AFPC/DSYX) initiated a contract to evaluate 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. The details of the analyses performed on the cognitive subtests and the SDI-O are the subject of several different technical reports. This technical report serves as the final deliverable for the research examining the SJT portion of the test only.

Although the performance work statement (PWS) did not specify any particular hypotheses, Infoscitex (IST) identified a series of research questions to evaluate the SJT potential for the future inclusion in the AFOQT operational composites. These research questions included: What are the SJT psychometric properties? What is the SJT Reading Grade Level? How does the SJT relate to the cognitive subtests and to the SDI-O? Are there subgroup differences on the SJT? Is there test-retest reliability? Are the SJT scores on T1 and T2 equivalent? Do the test-takers improve their SJT scores when retesting? Was there a breach of test security between earlier and later test administration dates? Do demographics variables explain variance in the SJT scores?

The item-level analyses included: Descriptive statistics, item difficulty and chance responding, item discriminability, internal consistency if item is deleted, distractor analysis, subgroup differences, item and doublet omissions, doublet longstring, item drift, analysis of common items, Principal Components Analysis (PCA), Reading Grade Level (RGL) analysis, doublet longstring rate, and an evaluation of several scoring methods.

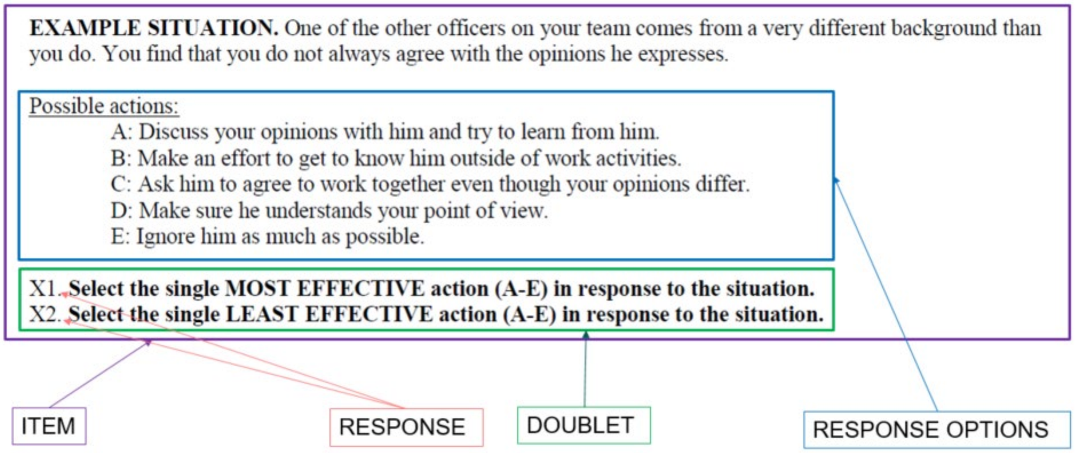
The subtest-level analyses included: Descriptive statistics, subtest difficulty, subtest discriminability, internal consistency (Cronbach's alpha [α; Cronbach, 1951] and McDonald's omega [ω; Flora, 2010]), subgroup differences, subtest correlations, scores by demographics, test-retest and alternate forms reliability, retesting effects (i.e., comparisons of those test-takers who retested once, twice, or three times), and stability analysis. These subtest-level analyses were completed for overall SJT scores, scores on ME responses only, scores on LE responses only, and scores on all six of the aforementioned competencies.

1.2.1 SJT Terminology

To facilitate reading clarity when describing the SJT analyses, we provide the following terminology (see Figure 1).

In the 'Item-Level Results' section, the term *item* refers to the entire SJT object including the situation, the five possible actions, and test-takers' selection of an ME response option and an LE response option. Therefore, there are 25 items on T1 and 25 items on T2. Of these 25 items, 13 items are common and 12 items are unique between T1 and T2. The term *situation* refers to the short passage at the start of each item that presents the test-taker with a situation to which they would be asked to respond. Therefore, there are 25 situations on T1 and 25 situations on T2. The term *response option(s)* refer to the possible actions from which test-takers are to select the ME and LE response. There are five response options attached to each situation. The term *responses* refers to the ME or LE selections made by test-takers. Therefore, there are 50 responses for each test-taker for T1 and 50 responses for each test-taker for T2 (two for each situation). The term *doublet* refers to a combination of the ME and LE responses to each situation. Therefore, there are 25 doublets for T1 and 25 doublets for T2. While the majority of the analyses were performed at the response-level, some analyses were performed at the item- and doublet-level. The authors made every effort to keep the terminology consistent throughout this section of the report. However, there are instances when the terminology defaults to 'items,' rather than a more finite level of responses or doublets, since the term 'items' is more inclusive. For all item-level analyses, the results are presented at either the response-level, doublet-level, or both.

In the 'Subtest-Level Results' section, the term *subtest* refers to an average score obtained for either the overall SJT (across all 50 responses), ME (across 25 ME responses), LE (across 25 LE responses), or the six competencies (across the responses that mapped to each competency).



EXAMPLE SITUATION. One of the other officers on your team comes from a very different background than you do. You find that you do not always agree with the opinions he expresses.

Possible actions:

A: Discuss your opinions with him and try to learn from him.
B: Make an effort to get to know him outside of work activities.
C: Ask him to agree to work together even though your opinions differ.
D: Make sure he understands your point of view.
E: Ignore him as much as possible.

X1. Select the single MOST EFFECTIVE action (A-E) in response to the situation.
X2. Select the single LEAST EFFECTIVE action (A-E) in response to the situation.

Figure 1. SJT Terminology

2.0 METHOD

The results presented in the report are for the unstratified sample, which includes the first-time test-takers, unless otherwise stated. The results for the stratified samples are presented in the appendices and represent test-takers from one of the seven accession sources: (1) OTS-CIV, (2) OTS-AD, (3) AECP, (4) USAFA, (5) ROTC, (6) ANG, and (7) AFRES. The following sections describe the sample characteristics, data scrubbing steps, and the SJT scoring methods.

2.1 Sample Characteristics

Test-takers were candidates either for commissioning into USAF officership or for classification into a rated career field (ABM, CSO, pilot, and RPA pilot).

Table 2 shows the unstratified sample demographic composition. When a demographic is reported as 'Unknown,' it indicates that a test-taker passively declined to respond or that their response was not legibly marked (i.e., did not clearly select a response). 'Declined to Respond' indicates that a test-taker actively declined to provide a response (i.e., selected "Decline to Respond").

The mean age of T1 test-takers was 23.80 (*SD* = 5.04). The mean age of T2 test-takers was 23.65 (*SD* = 5.06). Appendix B contains the demographic information stratified by accession source.

Table 2. Unstratified Sample Demographics

Sample Characteristic	T1		T2	
	N	%	*N*	%
Sex				
Male	26,557	74%	24,277	74%
Female	9,488	26%	8,613	26%
Unknown	33	0%	42	0%
Race*				
American Indian/Alaska Native	1,938	5%	1,812	6%
Asian	3,728	10%	3,309	10%
Black or African American	5,005	14%	4,574	14%
Native Hawaiian/Other Pacific Islander	918	3%	824	3%
White	27,840	77%	25,250	77%
Ethnicity				
Hispanic	5,208	14%	4,742	14%
Non-Hispanic	30,204	84%	27,500	84%
Unknown	666	2%	690	2%
Socio-Economic Status				
Much higher than average	1,747	5%	1,704	5%
Somewhat higher than average	10,196	28%	9,420	29%
Average	14,076	39%	12,540	38%
Somewhat lower than average	6,215	17%	5,594	17%
Much lower than average	2,112	6%	2,009	6%
Declined to respond	1,625	5%	1,574	5%
Unknown	107	0%	91	0%
Years of Education				
Completed 12	1,175	3%	1,251	4%
Completed 13	10,379	29%	9,988	30%
Completed 14	4,841	13%	4,314	13%
Completed 15	3,862	11%	3,368	10%
Completed 16	10,789	30%	9,439	29%
Completed 17	2,293	6%	2,068	6%
Completed 18	1,849	5%	1,697	5%
Completed 19	399	1%	359	1%
Completed 20	218	1%	204	1%
Completed 21+	221	1%	190	1%
Unknown	52	0%	54	0%
Highest Academic Degree				
High School Diploma	17,407	48%	16,353	50%
Associates Degree	3,741	10%	3,392	10%
Bachelor's Degree	13,021	36%	11,400	35%
Master's Degree	1,692	5%	1,628	5%
Doctoral Degree	128	0%	89	0%
Unknown	89	0%	70	0%
Accession Source				
OTS-CIV	8,022	22%	7,282	22%
OTS-AD	5,428	15%	4,837	15%
AECP	170	0%	168	1%
USAFA	4,303	12%	3,736	11%
ROTC	12,759	35%	12,406	38%
ANG	3,915	11%	3,288	10%
AFRES	1,481	4%	1,215	4%

*The proportions for Race do not add to 100% because test-takers had an option to choose more than one race.
Note. N = Sample Size. N_{T1} = 36,078 and N_{T2} = 32,932.

2.2 Technical Approach

Data were collected between 2014 and 2020 and were de-identified (i.e., personally identifiable information [PII] was removed). To track participants, unique alpha-numeric identifiers (i.e., SPARKIDs) were assigned to each test-taker. Data were then scrubbed as outlined below. Finally, the scrubbed data were subjected to various statistical analyses in accordance with Classical Test Theory (CTT). For a complete list of the generated statistics and their interpretation see Appendix C. Note that IST performed additional analyses, such as Generalizability Theory (G-Theory) and stepwise regression, which are not covered in the present technical report. For more detail, please refer to Appendix U.

2.2.1. Data Scrubbing

The 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).

IST received the scrubbed dataset from the AFPC/DSYX. The dataset included 78,397 cases across T1 and T2. Test-takers who self-identified as belonging to the Test Control Officer-Test Evaluator (TCO-TE), AD-Other (Active Duty-Other), or CIV-Other (Civilian-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, 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. These variables consisted largely of the individual item responses for the ten cognitive subtests and the SDI-O.

Finally, the dataset was scrubbed to address missingness (see Section 2.2.1.1) and to address carelessness, or Insufficient Effort Responding (IER; see Section 2.2.1.2). Cases that were identified as unit nonresponse and unit IER were removed listwise. Next, the cognitive subtest scores were imported. Note that the cognitive subtest scores for the cases that had unit nonresponse and unit IER on each of the cognitive subtests were replaced with Not Available (NA). Additional datasets were created to compare the psychometric properties of the SJT scored with three different scoring methods: (1) dichotomous, (2) trichotomous, and (3) continuous. Details for these scoring options are provided in Section 2.3.

2.2.1.1 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 SJT 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 not the responses on a subtest; Schafer & Graham, 2002). *Item-nonresponse* occurs when the information is available for some items but not for other items (e.g., a test-taker provided responses to 10 of 25 items on the SJT; Schafer & Graham, 2002).

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

There are three common terms associated with 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, refer to Schafer and Graham (2002). In brief, *MCAR* means that a missing response on variable X for a test-taker does not depend on their own values on variable X and another variable Y (and, by independence, does not depend on variable X and Y of other test-takers). *MAR* means that the missing response on Variable X may 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.

It was important to examine missingness also due to the fact that the SJT is not part of the AFOQT operational composites – which is common knowledge disseminated to test-takers via the AFOQT pamphlets. This means that the SJT may be considered low-stakes and test-takers' motivation to perform well may be low. Recall that test-taking motivation is defined as "willingness to engage as well as invest effort and persistence in working on test items" (Rios et al., 2017, p. 74). The examination of SJT missingness at the item-level revealed that the percentage of missing data increased throughout the test (i.e., a lower percentage of missing data in the first half of the test and a higher percentage of missing data in the second half of the test). However, the highest amount of missingness was less than 5%, which is a common rule of thumb in the literature for data missing MCAR (Schafer & Graham, 2002). This also suggests a possibility that test-takers were running out of time to complete the SJT (i.e., the SJT may be slightly speeded; see Section 4.7 for more details).

The examination of SJT missingness at the test-taker-level identified test-takers who did not provide any responses for the entire SJT (i.e., unit nonresponse); test-takers who omitted individual responses such as ME *or* LE (i.e., response omission); and test-takers who omitted both ME *and* LE (i.e., doublet omission). Several strategies were proposed to handle missingness. Ultimately, we decided to remove unit nonresponse listwise from the data set as this helps reduce noise in the data, but does not imply assumptions about motivation, nor does it potentially remove low ability test-takers. This strategy is conservative, and ensures that we do not remove potentially valid responses from the data set.

2.2.1.2 Carelessness

Low test-taking motivation may result in unit nonresponse (as described above) or in responses which do not reflect the test-taker's best effort, termed as *IER*. Potential impacts of IER include lower validity of score-based inferences, biased individual ability estimates, biased item parameter estimates, biased reliability estimates, and biased correlations with external variables (Huang et al., 2015). Given that the SJT is not part of the AFOQT operational composites, low test-taker' motivation could result in higher levels of IER.

Many methods can be used to detect IER, each of which can capture different types of IER with different levels of sensitivity (Huang et al., 2015; Meade & Craig, 2012). These methods include pre-test and post-test strategies. *Pre-test strategies* include instructed response items, infrequency items, and self-reported items (Meade & Craig, 2012). *Post-test strategies* can be organized into response-pattern-based methods, response-latency-based methods, and individual-reliability-based methods (Meade & Craig, 2012).

Since the present research utilized archival data, only post-test strategies were considered. Of the three aforementioned post-test strategies, the response latency-based method was not possible because the SJT (and the entire AFOQT Form T) is administered as a paper-and-pencil test. Thus, the other two methods were considered. Ultimately, we decided to only use longstring – a response-pattern based method that identifies straight-line response patterns (e.g., responding with "C" for all items in a row). This method is simple, matches the IER detection method we used for cognitive subtests, and captures the 'low-hanging fruit' or the worst and most obvious type of carelessness. We removed participants listwise if they engaged in 50-item-long longstring (i.e., unit-level longstring). We used this maximum level of longstring in order to err on the side of Type II errors, and avoid removing potentially valid responses from the data set.

After applying the aforementioned data scrubbing steps and removing unit careless responding and unit missingness, the initial dataset was split into several datasets to accommodate the various specialized analyses. For a step-by-step data scrubbing process, please refer to Attachment 1.

2.3 SJT Scoring

The literature suggests that the scoring method may impact the SJT psychometric properties, such as criterion-related validity, factor structure, subgroup differences, faking, internal consistency, and legal defensibility (Arthur et al., 2014; Bergman et al., 2006; Weng et al., 2018). Therefore, IST recommended exploring the most applicable scoring methods to identify the strongest method. Table 3 describes several well-researched scoring methods.

Table 3. Scoring Methods

Scoring Method	Description
Empirical	Options are scored based on their relationships with a criterion measure.
Theoretical	Options are scored based on their relationships with a theoretical construct or constructs.
Expert-based	Options are scored based on expert ratings. Expert ratings may be keyed dichotomously, trichotomously, or continuously.
Factorial	Options are scored based on factor analysis and item inter-correlations.
Subgrouping	Options are scored based on identifying groups of test-takers with similar patterns of responding.
Hybrid	Two independently generated scores are combined (e.g., using an empirical scoring method to eliminate the least predictive items, then use a theoretical scoring method for the remaining items).

Note. Based on Bergman et al. (2006).

2.3.1. Original Scoring

As previously noted, the original SJT scoring method was established by engaging military SMEs who provided effectiveness ratings for each response option using a Likert scale ranging from 1 (‘*Very ineffective action to address the situation’*) to 7 (‘*Very effective action to address the situation’;* Barron, 2013). To be keyed as the ME option, the response must have been judged as *more* effective than each of the other alternatives. To be keyed as the LE option, the response must have been judged as *less* effective than each of the other alternatives. While some items ended up having clear LE and ME keys (i.e., a large difference between the ME or LE response and the next best response option), others had more ambiguous ME and LE keys (i.e., a small or inconsequential difference between the ME or LE response and the next best response option).

Using the Motowidlo et al. (1990) scoring method, test-takers could gain a point (‘+1’) if their ME/LE responses matched those of the SMEs (i.e., a “correct” response option), lose a point (‘-1’) if their selection was a mismatch to the SME selection (i.e., they selected as ME the SME LE selection or vice versa; an “incorrect” response option), and neither gain nor lose a point (‘0’) if their ME/LE responses were neither a match, nor a mismatch to the SME selection (i.e., a “neutral” response option, neither ‘effective’ nor ‘ineffective’). For each item, there were any number of ME, LE, and neutral response options. For example, for Item 1 on the AFOQT Form T1, there were two ME options, two neutral options, and one LE option. For Item 9 on Form T1 and for Item 8 on Form T2 (this is a common item), there were no LE options, two neutral options and three ME options.

The Motowidlo et al. (1990) scoring method was replaced, such that ‘-1’ points were rescored as ‘1’, ‘0’ points were rescored as ‘2’, and ‘+1’ points were rescored as ‘3’. So, the final dataset that IST received from the AFPC/DSYX contained scores ranging from 1 to 3. However, a large proportion of SJT responses were annulled (i.e., removed from scoring consideration). The impetus for the annulment was a previously conducted criterion-related validity study in which some items were deemed criterion-invalid. To be deemed criterion valid, an item must have correlated at a minimum of .03 with at least two of three leadership outcomes (Air Force Field Training Relative Standing Score [RSS], Army Field Training, USAFA Military Performance Average [MPA]) in the largest available training sample (Army Field Training *N* = 3,587) and must have not correlated negatively with RSS. A correlation of .03 was chosen because it represented statistical significance ($p \leq .05$) in that sample. Please refer to Attachments 2 and 3.

Following the item annulment, the dataset contained the subtest composite score (i.e., SJT_SCO) that included only cases in which test-takers had responded to at least 70% of the 24 criterion-valid items (i.e., at least 17 items). Test-taker responses were averaged across the valid items to create this composite score. This procedure limited the knowledge that could have been gleaned from analyzing the entire dataset. Therefore, IST recommended that alterative scoring methods be explored using the entire item bank (i.e., all 50 responses for Form T1 and 50 responses for Form T2). This recommendation was supported by the AFPC/DSYX.

2.3.2. Alternative Expert-Based Scoring Methods

Of the scoring methods described by Bergman et al. (2006), IST recommended exploring the expert-based method (which would expand on the Barron [2013] strategy) and the empirical method, not previously explored. To generate the expert-based scoring method, the mean SME effectiveness ratings collected as part of the Barron study were utilized to generate dichotomous, trichotomous, and continuous scoring keys.

As previously mentioned, Barron (2013) used a scale from 1 (‘*Very Ineffective action to address the situation’*) to 7 (‘*Very Effective action to address the situation*’) to generate mean effectiveness ratings. The next few paragraphs describe how these effectiveness ratings were used to generate each scoring method and its corresponding cutoffs for the ME, LE, and neutral responses.

Dichotomous
In the dichotomous scoring method, correct responses were keyed as ‘1’ and incorrect or neutral responses were keyed as ‘0.’ For the ME scoring key, the SME mean effectiveness ratings of 5 and above were deemed as correct. Mean effectiveness ratings between 4 and 4.99 were deemed as neutral. Mean effectiveness ratings of below 4 were deemed as incorrect. The reverse was true for the LE scoring key. Thus, it was possible for more than one response option to be keyed as correct for the ME or LE response or for a response to have no correct response option.

Trichotomous
In the trichotomous scoring method, correct responses were keyed as ‘3,’ incorrect responses were keyed as ‘1’, and neutral responses were keyed as ‘2.’ For the ME scoring key, the SME mean effectiveness ratings of 5 and above were deemed as correct. Mean effectiveness ratings between 4 and 4.99 were deemed as neutral. Mean effectiveness ratings of below 4 were deemed as incorrect. The reverse was true for the LE scoring key. Thus, it was possible for more than one response option to be keyed as correct for the ME or LE response or for a response to have no correct response option. Of note, this is the method currently employed in scoring the AFOQT Form T SJT.

Continuous
In the continuous scoring method, all response options were rank-ordered according to the SME mean effectiveness ratings (in ascending order for the ME and in descending order for LE). For the ME scoring key, the highest rating was deemed as correct and the lowest as incorrect. The

reverse was true for the LE scoring key. Thus, every response had exactly one correct response option, one incorrect response option, and three neutral response options.

See Table 4 for an example of the three expert-based scoring methods described above. See Appendix D for the full scoring keys.

Table 4. Examples of the Three Expert-Based Scoring Methods

Response Options	SME *M*	Dichotomous		Trichotomous		Continuous	
		ME	LE	ME	LE	ME	LE
A	1.55	0	1	1	3	1	5
B	4.93	0	0	2	2	3	3
C	4.23	0	0	2	2	2	4
D	5.72	1	0	3	1	5	1
E	5.39	1	0	3	1	4	2

Note. SME = Subject Matter Expert; *M* = Mean Effectiveness Rating; ME = Most Effective; LE = Least Effective.

Items with Single vs. Multiple Correct Response Options

As previously mentioned, some responses had a clear correct or incorrect mean effectiveness rating, while others had multiple correct or incorrect effectiveness ratings. This issue has advantages and disadvantages. Among the advantages is the fact that allowing more than one correct response option per response resolves the issue of tied SME ratings in the dichotomous and trichotomous keys (i.e., the ratings that are too close to one another to be meaningfully distinctive). Among the disadvantages is the fact that the calculation of test-takers' ability may be complicated[1], because each item would by definition have varying levels of difficulty based on random chance of selecting the correct response. Table 5 shows an example of an item (Item 3) for which test-takers would have a 1/5 (i.e., 20%) chance of endorsing the correct ME response option (i.e., 'A') and the same (i.e., 20%) chance of endorsing the correct LE response option (i.e., 'E'). Conversely, on Item 1, test-takers would have a 2/5 (i.e., 40%) chance of endorsing a correct ME response option (i.e., 'D' or 'E') and 1/5 (i.e., 20%) chance of endorsing the correct LE response option (i.e., 'A').

[1] In theory, ME and LE responses should be locally dependent (i.e., test-takers should endorse the ME response first and then they should eliminate that response option as a possible response to the LE item) which would mean chance responding would be different for the ME and LE responses due to a different number of response options available to select. Because the SJT is a paper-and-pencil test, there is no way to control for order of responding (e.g., participant may endorse the LE item prior to the ME item). Additionally, there is nothing precluding test-takers from selecting the same response option for both the ME and LE response. Due to inability to accurately determine chance responding, responses were treated as locally independent and chance responding was set at 20% per correct response for both ME and LE items for all SJT situations

Table 5. Comparing Items with Single and Multiple Correct Responses

Item	Response Options	Mean SME Effectiveness Ratings	Keyed for ME	Keyed for LE
T1 Item 3 (single correct response)	A	5.17	Correct	Incorrect
	B	4.18	Neutral	Neutral
	C	4.87	Neutral	Neutral
	D	4.91	Neutral	Neutral
	E	1.89	Incorrect	Correct
T1 Item 1 (multiple correct responses)	A	1.55	Incorrect	Correct
	B	4.93	Neutral	Neutral
	C	4.23	Neutral	Neutral
	D	5.72	Correct	Incorrect
	E	5.39	Correct	Incorrect

Note. SME = Subject Matter Expert; ME = Most Effective; LE = Least Effective.

2.3.3. Alternative Empirical Scoring Method

The final method involved keying the items based on their empirical relationship with criteria. The raw response options for each item were dummy coded (e.g., for the first dummy code, if a test-taker endorsed A, they received a 1, all else 0; for the second dummy code if a test-taker endorsed B, they received a 1, all else 0). Thus, each item had five dummy codes, one per response option. These dummy coded variables were then correlated with important criteria.

There are two steps involved in generating an empirically-derived scoring key (Bergman et al., 2006). The first step is to determine the option endorsement rate and the second step is to determine the option-criterion correlation threshold below which the correlations are deemed low. The AFPC/DSYX provided IST with a list of 26 different officership criteria. Of these, 16 were deemed as most applicable to the SJT competencies (e.g., Peer Ranking Academic Achievement, Final Score). The decision rules established were as follows: The option endorsement rate needed to be at least 20% of the sample and the option-criterion correlation needed to be at least .10.

Correlations between all response options and the 16 criteria were examined. It was determined that the resulting keys were not robust. For example, the same response option was sometimes correct for both the ME and LE item (which should not be the case) and keys for common items between forms T1 and T2 were different. Based on these inconsistencies, the decision was made to abandon this scoring method in favor of the expert-based scoring methods. Therefore, all comparative analyses included only the three expert-based scoring methods: (1) dichotomous, (2) trichotomous, and (3) continuous.

2.3.4. Comparing Scoring Methods

The three expert-based scoring methods were compared on several item-level psychometric properties, such as descriptives, mean score subgroup differences, and correlations with the aforementioned SJT-relevant criteria. The differences among the three scoring methods were generally small with no clear indication as to which scoring method was superior.

Given these findings, IST recommended the trichotomous scoring method based on the following considerations:

- Dichotomous scoring keys may confuse the experts' consensus on the most/least effective response option with a correct response. SJTs do not have an objectively correct/incorrect response (in contrast to cognitive ability tests; Bergman et al., 2006). Sometimes even experts disagree.
- Continuous scoring keys assign different score values to response options that are objectively very similar (e.g., compare a SME effectiveness rating of 5.14 with 5.15). Thus, scores become somewhat arbitrary for these items with inconsequential differences between ME and LE options and the next best option.
- The trichotomous scoring rubric is used for the current SJT and barring conclusive evidence of a better scoring option, there is no evidence-based reason to change the scoring methods. Continuing to use the same scoring method makes the interpretation of the existing scores easier and more intuitive and comparison of scores across forms simpler.

Discussion with the AFPC/DSYX resulted in the adoption of the trichotomous scoring option. Based on this decision, the remainder of this report summarizes the results using the trichotomous scoring method only. It is recommend that future efforts attempt to replicate these results and explore other scoring methods.

3.0 ITEM-LEVEL RESULTS

This section covers the item-level (doublet- and response-level) analyses and results. The results reported in the text will be for the unstratified sample. Results stratified by accession sources are provided in the appendices.

3.1 Item Descriptives

Table 6 displays the results of the descriptive analyses. As can be seen, the responses were relatively easy with only moderate variance (T1: *M* = 2.66, *SD* = .31; T2: *M* = 2.67, *SD* = .29). The responses were generally negatively skewed which means that more people selected the correct responses than what would be expected if the responses were normally distributed (T1: $M_{Skewness}$ = -2.73; T2: $M_{Skewness}$ = -2.51). Responses 16 and 28 on T1 and Response 12 on T2 demonstrate extreme kurtosis due to significant lack of variance (i.e., almost all test-takers selected the correct response). For stratified results, see Appendix E.

Table 6. Unstratified Sample Item-Level Descriptives

Item	Response	T1					T2				
		N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
1	1	36,052	2.93	.26	-3.94	15.47	32,902	2.80	.40	-1.59	.77
	2	36,068	2.95	.27	-5.86	35.21	32,915	2.94	.26	-4.19	18.03
2	3	36,034	2.82	.43	-2.34	4.90	32,870	2.19	.83	-.36	-1.46
	4	36,048	2.94	.30	-5.64	31.58	32,893	2.91	.38	-4.43	18.58
3	5	36,032	2.78	.42	-1.47	.48	32,898	2.77	.52	-2.26	4.15
	6	36,050	2.94	.26	-4.25	18.79	32,898	2.33	.75	-.62	-.96
4	7	36,026	2.27	.78	-.52	-1.17	32,890	2.41	.88	-.90	-1.10
	8	36,039	2.93	.34	-4.94	23.73	32,911	2.63	.60	-1.37	.81
5	9	36,037	2.79	.50	-2.40	4.92	32,890	2.38	.60	-.40	-.67
	10	36,047	2.36	.73	-.69	-.86	32,912	2.84	.45	-2.82	7.30
6	11	36,053	2.82	.42	-2.15	3.87	32,893	2.35	.94	-.76	-1.43
	12	36,063	2.89	.32	-2.80	6.80	32,911	2.99	.17	-11.61	132.70
7	13	36,036	2.35	.94	-.75	-1.44	32,896	2.87	.50	-3.48	10.13
	14	36,042	1.99	1.00	.02	-2.00	32,910	2.93	.36	-5.14	24.44
8	15	36,041	2.45	.89	-1.01	-.97	32,883	2.83	.37	-1.78	1.18
	16	36,051	2.99	.17	-11.67	134.19	32,906	1.56	.50	-.25	-1.94
9	17	36,024	2.80	.40	-1.51	.29	32,897	2.82	.40	-1.94	2.61
	18	36,040	1.59	.49	-.36	-1.87	32,909	2.50	.70	-1.06	-.22
10	19	36,035	2.27	.88	-.56	-1.48	32,888	2.84	.38	-2.19	3.79
	20	36,057	2.92	.32	-4.25	18.70	32,907	2.86	.47	-3.24	9.31
11	21	36,019	2.64	.77	-1.67	.80	32,891	2.49	.51	-.05	-1.76
	22	36,034	2.94	.35	-5.36	26.78	32,900	2.69	.55	-1.62	1.68
12	23	36,032	2.85	.38	-2.41	5.05	32,886	2.95	.28	-6.30	39.40
	24	36,037	2.85	.49	-3.18	8.71	32,894	2.79	.48	-2.32	4.65
13	25	36,010	2.26	.94	-.53	-1.65	32,845	2.20	.95	-.41	-1.77
	26	36,004	2.89	.36	-3.58	12.87	32,858	2.89	.37	-3.42	11.68
14	27	36,009	2.45	.80	-.97	-.75	32,849	2.86	.37	-2.65	6.52
	28	36,030	2.99	.15	-12.19	153.62	32,859	2.75	.64	-2.24	3.18
15	29	35,995	2.48	.84	-1.08	-.73	32,778	2.21	.65	-.24	-.71
	30	35,988	2.96	.23	-6.03	39.11	32,794	2.25	.66	-.32	-.76
16	31	35,938	2.81	.42	-2.09	3.54	32,778	2.77	.51	-2.12	3.63
	32	35,957	2.65	.73	-1.73	1.12	32,781	2.75	.46	-1.58	1.44
17	33	35,914	2.82	.46	-2.60	6.06	32,703	2.83	.42	-2.50	5.75
	34	35,928	2.92	.31	-3.97	16.44	32,729	2.96	.26	-7.00	48.33
18	35	35,841	2.86	.52	-3.31	8.94	32,631	2.79	.50	-2.37	4.74
	36	35,842	2.79	.62	-2.54	4.44	32,629	2.91	.33	-3.83	15.06
19	37	35,702	2.42	.83	-.91	-.92	32,474	2.18	.84	-.36	-1.49
	38	35,692	2.36	.82	-.75	-1.09	32,487	2.95	.30	-5.73	32.25
20	39	35,562	2.74	.57	-2.12	3.28	32,264	2.42	.84	-.90	-.96
	40	35,572	2.85	.50	-3.15	8.46	32,265	2.36	.79	-.73	-1.01
21	41	35,381	2.64	.52	-.97	-.23	32,110	2.92	.34	-4.60	20.91
	42	35,340	2.45	.71	-.90	-.52	32,097	2.62	.73	-1.59	.73
22	43	35,173	2.92	.33	-4.63	21.54	31,854	2.79	.62	-2.55	4.51
	44	35,148	2.62	.74	-1.56	.65	31,849	2.86	.51	-3.39	9.52

23	45	34,889	2.78	.63	-2.49	4.19	31,485	2.76	.59	-2.32	3.87
	46	34,867	2.86	.51	-3.35	9.23	31,521	2.80	.58	-2.63	5.18
24	47	34,609	2.05	.92	-.09	-1.82	31,131	2.65	.69	-1.67	1.15
	48	34,553	2.01	.93	-.02	-1.85	31,123	2.86	.45	-3.32	10.05
25	49	34,279	2.78	.57	-2.43	4.42	31,012	2.32	.95	-.68	-1.54
	50	34,312	2.80	.57	-2.70	5.56	30,968	2.94	.33	-5.66	30.06

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

3.2 Item Difficulty

Difficulty, expressed as a probability value (*p*-value), refers to the proportion of test-takers who endorsed the correct response (Kline, 2005). The higher the *p*-value, the easier the item, because more test-takers endorsed the correct response. The reverse is true of the lower *p*-values. The lower the *p*-value, the harder the item, because fewer test-takers endorsed the correct response.
Table 7 displays the results of the difficulty analyses. Similar to the SJT response means, the difficulty parameters also suggest that the SJT is relatively easy with an average *p*-value of .78 for Form T1 and .76 for Form T2. This indicates that on average 78% and 76% of test-takers endorsed correct responses for the entire T1 and T2 SJT, respectively. A few responses on each form were particularly problematic in that nearly every test-taker selected the correct response (i.e., T1: Response 16 and 28; Form T2 Response 12). Additionally, two responses on each form were also problematic in that fewer participants selected the correct response than would be expected due to random chance, suggesting a problematic distractor (i.e., T1: Responses 47 and 48; Form T2 Responses 29 and 30).

3.2.1. Item Difficulty and Chance Responding

As noted previously, the chance of randomly selecting the correct response option (or *chance responding*) is different for each ME and LE response on the SJT. Because of this, a response with a 60% chance of selecting the correct response options should be inherently easier than a response with a 20% chance of selecting the correct response option. Therefore, the overall difficulty of an item should be considered by examining both the *p*-value and the chance of randomly selecting the correct response.

Table 7 also displays the product of the difficulty and the chance responding for each doublet. To calculate these statistics, the following steps were undertaken. First, each response's *p*-value was multiplied by its chance responding percentage. For example, for the first response on T1, the *p*-value was .94 and chance responding was 40% (i.e., there were two correct response options). Thus, the overall difficulty was .37. For the second response on T2, the *p*-value was .96 and chance responding was 20% (i.e., there was one correct response option). Thus, the overall difficulty was .19. Second, these two values were averaged to derive the doublet-level difficulty. The average between .37 and .19 is .28.

Considering *p*-values and chance responding in combination *and* separately can help identifying overly easy or overly difficult items and can possible help pinpointing the reasons for why that is (e.g., poor distractors, too many correct options). The current results revealed that the proportion of test-takers who endorsed the correct ME response option was similar to the proportion of test-takers who endorsed the correct LE response option, with the LE response being slightly easier (averaged doublet difficulty: T1: M_{ME} = .30, M_{LE} = .34; T2: M_{ME} = .28, M_{LE} = .33). This is consistent with the theory of Implicit Trait Policy (ITP) by Motowidlo and Beier (2010). ITP suggests that it may be easier for individuals to know (or guess) what the least effective response is through the process of fundamental socialization (i.e., it is easier to determine socially inappropriate behavior than socially appropriate behavior). This taps into non-job-specific knowledge. Conversely, individuals may be less likely to know what behaviors are most effective, especially in the military setting. This taps into job-specific knowledge. For the stratified samples item difficulty, see Appendix F.

Table 7. Unstratified Sample Item-Level Difficulty

Item	Response	*p*-value		Doublet Difficulty		Item	Response	*p*-value		Doublet Difficulty	
		T1	T2	T1	T2			T1	T2	T1	T2
1	1	.94	.80	.28	.17	14	27	.64	.87	.36	.35
	2	.96	.94				28	.99	.86		
2	3	.84	.46	.36	.33	15	29	.71	.34	.36	.07
	4	.96	.94				30	.96	.37		
3	5	.79	.82	.17	.21	16	31	.83	.81	.33	.23
	6	.94	.50				32	.81	.77		
4	7	.48	.67	.33	.27	17	33	.85	.85	.27	.37
	8	.95	.69				34	.93	.98		
5	9	.84	.44	.22	.22	18	35	.93	.83	.46	.27
	10	.52	.87				36	.89	.92		
6	11	.83	.68	.26	.43	19	37	.64	.46	.24	.34
	12	.90	.99				38	.58	.97		
7	13	.67	.93	.30	.48	20	39	.81	.64	.34	.24
	14	.50	.97				40	.90	.55		
8	15	.73	.83	.44	.25	21	41	.66	.94	.19	.36
	16	.99	NA				42	.58	.78		
9	17	.80	.83	.24	.17	22	43	.94	.89	.36	.46
	18	NA	.62				44	.77	.93		
10	19	.56	.85	.34	.34	23	45	.89	.85	.46	.35
	20	.93	.90				46	.93	.88		
11	21	.82	.49	.45	.17	24	47	.45	.77	.18	.35
	22	.97	.74				48	.44	.90		
12	23	.86	.97	.35	.37	25	49	.85	.66	.35	.45
	24	.91	.83				50	.89	.97		
13	25	.60	.58	.33	.33						
	26	.91	.90								

Note. NA indicates that the parameters could not be estimated, because these items did not have a correct LE response option.

3.3 Item Discriminability

Discriminability, expressed as an item-total correlation (ITC), refers to a correlation between each item and the total scale score (computed with the item in question removed; Kline, 2005). Much like the Pearson product moment correlation, ITCs range from -1 to 1, where high positive values indicate a strong association between the item and the entire scale; low positive values indicate a weak association between the item and the entire scale; and negative values indicate that the item is negatively related to the entire scale, which is highly undesirable.

Table 8 displays the results of the discriminability analyses. Of note, the lowest ITC on both forms were negative, which may suggest that the response needs to be dropped (or perhaps the entire item needs to be revised). Overall, the ITC's tend to be low to moderate, suggesting the test has low discriminability. Discriminability statistics for the stratified samples are available in Appendix G.

Table 8. Unstratified Sample Item-Level Discriminability

Item	Response	ITC		Item	Response	ITC	
		T1	T2			T1	T2
1	1	.09	.11	14	27	.11	.09
	2	.11	.14		28	.12	.15
2	3	.07	.18	15	29	.12	.10
	4	.12	.15		30	.10	.11
3	5	.09	.23	16	31	.08	.20
	6	.14	.19		32	.12	.11
4	7	.17	-.02	17	33	.20	.12
	8	.13	.09		34	.18	.24
5	9	.19	.14	18	35	.19	.26
	10	.16	.13		36	.24	.25
6	11	.13	.07	19	37	.17	.11
	12	.11	.12		38	.11	.21
7	13	.11	.20	20	39	.13	.17
	14	.12	.15		40	.20	.12
8	15	.09	.12	21	41	.15	.24
	16	.13	.07		42	.22	.28
9	17	.11	.09	22	43	.22	.27
	18	.05	.16		44	.22	.25
10	19	.09	.10	23	45	.23	.32
	20	.09	.16		46	.23	.28
11	21	.11	.07	24	47	-.04	.30
	22	.12	.13		48	.00	.27
12	23	.13	.12	25	49	.24	.23
	24	.16	.16		50	.23	.17
13	25	.10	.09				
	26	.10	.10				

3.4 Internal Consistency if Item is Deleted

Internal consistency, expressed as Cronbach's alpha, concerns the interrelatedness of items (Schmitt, 1996). At the response level, internal consistency shows how the alpha would change if the response is deleted. Table 9 displays the Cronbach's alphas. Overall, the results suggest that internal consistency could not be improved drastically by removing some problematic responses (the largest change in alpha would be an increase by .03 on both Forms T1 and T2). Refer to Section 4.5 for a detailed discussion of the subtest-level Cronbach's alpha and McDonald's omega. For stratified samples reliability statistics, see Appendix H.

Table 9. Unstratified Sample Item-Level Internal Consistency if Item is Deleted

Item	Response	α		Item	Response	α	
		T1	T2			T1	T2
1	1	.56	.63	14	27	.56	.63
	2	.56	.63		28	.56	.63
2	3	.56	.63	15	29	.56	.63
	4	.56	.63		30	.56	.63
3	5	.56	.63	16	31	.56	.63
	6	.56	.63		32	.56	.63
4	7	.55	.65	17	33	.55	.63
	8	.56	.63		34	.56	.63
5	9	.55	.63	18	35	.55	.63
	10	.55	.63		36	.55	.63
6	11	.56	.64	19	37	.55	.64
	12	.56	.63		38	.56	.63
7	13	.56	.63	20	39	.56	.63
	14	.56	.63		40	.55	.63
8	15	.56	.63	21	41	.56	.63
	16	.56	.64		42	.55	.62
9	17	.56	.63	22	43	.56	.62
	18	.56	.63		44	.55	.62
10	19	.56	.63	23	45	.55	.62
	20	.56	.63		46	.55	.62
11	21	.56	.64	24	47	.58	.62
	22	.56	.63		48	.58	.63
12	23	.56	.63	25	49	.55	.62
	24	.56	.63		50	.55	.63
13	25	.56	.64				
	26	.56	.63				

Note. α = Cronbach's alpha.

3.5 Item Distractor Analysis

Distractors refer to incorrect response options meant to attract low-ability test-takers (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 (i.e., is never selected). Distractor analysis was complicated because many responses had multiple correct answers and some responses had no correct options. Thus, the number of distractors varied from response to response. There were some distractors that likely warranted revision (i.e., no test-takers selected the response option, test-takers selected the response option at greater rates than at least one correct option), but given other psychometric problems with the items themselves, it is inappropriate to draw conclusions about whether distractors are ineffective because they are truly ineffective or because the item as a whole was ineffective.

3.6 Item Mean Score Subgroup Differences

One advantage of non-cognitive subtests such as the SJT is they have the potential to greatly reduce mean score subgroup differences over their cognitive counterparts (Nguyen et al., 2005; Whetzel et al., 2008). *Subgroup differences* refer to mean score differences between the majority group (e.g., White, males) and legally protected minority groups (e.g., racial and ethnic minorities, females). These differences are expressed using Cohen's *d*, which is a measure of mean score difference in standard deviation units (Cohen, 1992). Cutoffs of .40 and .80 were chosen to represent moderately and highly problematic subgroup differences, respectively. Although these cutoffs differ slightly from the ones prescribed by the literature, they are in line with the cutoffs used for the Armed Services Vocational Aptitude Battery (ASVAB) and other standardized assessments throughout the Department of Defense (DoD). Moderate to large effect sizes suggest that an SJT item (including the situation and the response options) may inadvertently favor one group over the other. The subgroups examined in this effort were – Female/Male (F/M), Black/White (B/W), Asian/White (A/W), White Non-Hispanic/White Hispanic (WnH/WH), and Black Non-Hispanic/Black Hispanic (BnH/BH).

Table 10 displays the results of the mean score subgroup differences. Consistent with the literature, Cohen's *d* on the SJT were small to moderate ranging from .05 to .17 for Form T1 and from .04 to .18 on Form T2. As expected, the effect sizes were generally smaller than those of the cognitive subtests (especially for F/M and B/W categories). Although the item-level subgroup differences were acceptable in the unstratified sample, there were larger subgroup differences in the stratified samples. See Appendix I for the item-level subgroup difference analyses for the stratified samples.

Table 10. Unstratified Sample Item-Level Subgroup Differences

Item	Response	T1 Cohen's *d**					T2 Cohen's *d**				
		F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
1	1	.07	.01	.04	.01	.06	.00	.11	.01	.04	.04
	2	.02	.09	.11	.05	.04	.01	.10	.03	.05	.02
2	3	.10	.12	.09	.06	.11	.03	.21	.03	.09	.03
	4	.01	.20	.08	.08	.01	.01	.10	.04	.05	.01
3	5	.02	.07	.05	.04	.06	.01	.16	.17	.10	.04
	6	.02	.05	.05	.03	.02	.02	.14	.18	.10	.07
4	7	.04	.21	.05	.08	.03	.02	.11	.02	.01	.15
	8	.02	.10	.03	.03	.10	.04	.20	.13	.08	.09
5	9	.00	.15	.16	.06	.07	.25	.17	.06	.08	.02
	10	.02	.12	.21	.04	.09	.05	.15	.07	.07	.01
6	11	.01	.13	.03	.02	.04	.01	.09	.09	.03	.02
	12	.02	.05	.04	.01	.00	.01	.08	.00	.04	.05
7	13	.10	.13	.16	.07	.01	.03	.23	.24	.10	.01
	14	.17	.13	.11	.05	.03	.01	.14	.09	.08	.06
8	15	.01	.10	.09	.07	.05	.04	.11	.19	.09	.03
	16	.00	.08	.04	.08	.02	.01	.06	.17	.05	.01
9	17	.02	.10	.17	.06	.07	.02	.04	.07	.00	.03
	18	.02	.06	.22	.07	.05	.10	.16	.32	.15	.01
10	19	.13	.06	.02	.03	.10	.12	.22	.15	.07	.12
	20	.01	.02	.06	.02	.03	.02	.09	.05	.03	.00
11	21	.00	.26	.19	.08	.02	.03	.13	.06	.04	.03
	22	.03	.14	.18	.06	.03	.02	.06	.15	.02	.02
12	23	.10	.28	.15	.10	.07	.03	.02	.05	.01	.05
	24	.03	.11	.03	.04	.00	.04	.08	.03	.01	.01
13	25	.01	.13	.02	.01	.02	.02	.14	.04	.03	.03
	26	.02	.10	.08	.04	.01	.01	.14	.03	.02	.08
14	27	.10	.10	.08	.07	.12	.04	.08	.12	.07	.04
	28	.00	.08	.03	.02	.02	.10	.10	.11	.10	.02
15	29	.07	.21	.11	.05	.01	.01	.07	.06	.06	.02
	30	.02	.08	.07	.00	.04	.10	.04	.11	.05	.14
16	31	.03	.10	.11	.03	.03	.03	.24	.04	.10	.04
	32	.14	.11	.12	.02	.01	.08	.02	.09	.00	.03
17	33	.03	.22	.19	.08	.02	.03	.25	.09	.11	.02
	34	.04	.17	.16	.06	.01	.01	.24	.06	.06	.04
18	35	.11	.32	.27	.13	.02	.02	.27	.21	.08	.02
	36	.11	.28	.19	.10	.05	.04	.28	.14	.14	.00
19	37	.10	.41	.18	.09	.09	.13	.08	.03	.07	.07
	38	.10	.22	.07	.04	.03	.03	.19	.01	.10	.03
20	39	.10	.12	.18	.06	.10	.08	.42	.16	.13	.01
	40	.03	.23	.12	.10	.15	.06	.25	.09	.08	.04
21	41	.05	.27	.11	.10	.04	.01	.25	.11	.14	.08
	42	.11	.31	.23	.14	.04	.01	.33	.27	.13	.07
22	43	.02	.26	.13	.09	.06	.03	.29	.26	.14	.02
	44	.02	.32	.27	.10	.02	.02	.29	.21	.12	.04

23	45	.04	.30	.32	.15	.05	.05	.40	.20	.18	.05
	46	.00	.25	.20	.09	.00	.08	.36	.20	.17	.03
24	47	.04	.15	.09	.09	.01	.07	.39	.15	.19	.01
	48	.03	.14	.12	.10	.07	.04	.26	.03	.14	.07
25	49	.02	.34	.16	.13	.00	.03	.28	.15	.11	.01
	50	.06	.33	.16	.14	.04	.03	.20	.06	.09	.03

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.
Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic.

3.7 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 expressed as an effect size (i.e., Cohen's *d*). Similar to subgroup differences, item drift parameters greater than .40 and greater than .80 may suggest moderate to large breach in test security.

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 latter group (test-takers who took the AFOQT between May 2018 and Dec 2019). The examination of the effect sizes comparing the early and the latter group revealed no evidence for item-level test security breach.

3.8 Common Items

As mentioned earlier, there are 13 common items between Forms T1 and T2 (i.e., identical situations and response options). These items should have similar psychometric properties. If there are differences, it may indicate that the two forms are not parallel, which may require equating, or it might 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.

Table 11 displays the results of the common items analysis. Overall, the results did not reveal any substantial differences between the two forms with *p*-values ranging from .0 to .05. See Appendix J for the common items analysis for the stratified samples.

Table 11. Unstratified Sample Common Items Analysis

Item		Response		*p*-value		*p*-value difference
T1	T2	T1	T2	T1	T2	(T1-T2)
3	1	5	1	.79	.80	.02
		6	2	.94	.94	.00
4	2	7	3	.48	.46	.02
		8	4	.95	.94	.01
5	3	9	5	.84	.82	.01
		10	6	.52	.50	.02
8	6	15	11	.73	.68	.05
		16	12	.99	.99	.00
9	8	17	15	.80	.83	.03
		18	16	NA	NA	NA
12	10	23	19	.86	.85	.01
		24	20	.91	.90	.00
13	13	25	25	.60	.58	.02
		26	26	.91	.90	.01
16	14	31	27	.83	.87	.05
		32	28	.81	.86	.05
17	18	33	35	.85	.83	.02
		34	36	.93	.92	.01
19	20	37	39	.64	.64	.00
		38	40	.58	.55	.03
22	21	43	41	.94	.94	.00
		44	42	.77	.78	.00
23	22	45	43	.89	.89	.00
		46	44	.93	.93	.00
25	23	49	45	.85	.85	.01
		50	46	.89	.88	.01

Note. NA indicates that the parameters could not be estimated, because these items did not have a correct LE response option.

3.9 Principal Components Analysis

The SJT purports to measure six competencies (described earlier) critical to officer performance. If these competencies are cleanly measured, a PCA should demonstrate a six-factor solution. This solution or any other interpretable solution would inform the creation of the subtest scores for the SJT.

Therefore, a Principal Components Analysis (PCA) including all 50 responses for T1 (and then for T2) was done which included only 25 responses for T1 ME (and then for T2 ME), and finally included only 25 responses for T1 LE (and then for T2 LE). The decision to run the PCA for ME and LE separately was based on the ITP theory (Motowidlo & Beier, 2010), according to which ME and LE may have different antecedents and outcomes. Specifically, it is possible that individuals need job-specific knowledge and cognitive ability to endorse a most effective strategy.

Conversely, individuals may need personality and general life knowledge to pick the least effective strategy. Therefore, we expected that these responses may result in different factor structures.

The factor structure of the SJT in its current form was uninterpretable and did not reveal clear factors. This is likely because SJTs are usually factorial complex and multidimensional (McDaniel et al., 2016; Ployhart & MacKenzie, 2011). The only clear finding is that the current SJT is multidimensional and may include competencies other than the six previously identified competencies.

PCAs were also conducted on the dichotomously and continuously scored data to see if the factor structure would be improved by using a different scoring method. However, no clear patterns emerged. Since no clear empirical solution emerged, a rational approach was used in creating subtest scores for the SJT (see Section 2.3 for more detail).

3.10 Reading Grade Level

Reading Grade Level (RGL) assesses how easy or difficult a selection of text is to comprehend. The Flesch-Kincaid formula is commonly used to calculate RGL. This method assesses average word and sentence length. Longer words and longer sentences indicate more difficult words, and more difficult syntax, which in turn denote a higher difficulty and RGL (see Figure 2 for Flesch-Kincaid formula). Note that there are multiple ways to calculate RGL for an SJT item. One way is to calculate RGL for the situation and the response options separately. Another method is to combine the situations with the response options and perform calculations. Note that IST performed both methods for comparison sake. Although fluctuating, both methods yielded comparable results.

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

Figure 2. Flesch-Kincaid Formula

Generally, an eighth-grade RGL (pre-high-school) is the target for employment assessments to ensure that items are understood by approximately 80% of American adults (Kutner et al., 2007). The overall SJT RGL was slightly higher than this target (T1: RGL = 9.47; T2: RGL = 9.95). The RGL for situations was slightly higher than the RGL for the response options (T1: $\text{RGL}_{\text{Situation}}$ = 10.08, $\text{RGL}_{\text{Response Options}}$ = 8.86; T2: $\text{RGL}_{\text{Situation}}$ = 10.72, $\text{RGL}_{\text{Response Options}}$ = 9.19). However, considering the academic degrees of the AFOQT test-takers, the majority of the sample have high school diploma (T1: 48%; T2: 50%) with the second largest proportion having Bachelor's degree (T1: 36%; T2: 35%), therefore, a slightly higher RGL may be appropriate. Additionally, the AFOQT test-takers represent officer and rated career field training candidates who need to have reading comprehension skills of at least a high-school grade level to understand and process the reading material that they will encounter in their training and careers. Therefore, it was recommended to reconsider the RGL requirements for the SJT to align with these job requirements. It is conceivable that the RGL may drive subgroup differences by the virtue of its heavy reliance on the knowledge of the English language and the common culture. This could be particularly problematic for test-takers for whom English is not their first language. Therefore, correlations were run between the RGL and the *p*-values at the doublet level. The working hypothesis was that as the RGL increases, the SJT difficulty would increase. For example, Item X with an RGL of 8 would have a higher average *p*-value (would be easier) than Item Y with an RGL of 9 (would be harder). Ultimately, we did not find support for this hypothesis. Correlations between the entire SJT RGL and the mean *p*-values were run for Forms T1 and T2. Note that since the higher *p*-values indicate that more test-takers endorsed the correct response, we used the term *easiness* instead of difficulty to clarify the directionality. The correlations (represented by *r*) were positive but non-significant (T1: *r* = .12, *ns*, T2: *r* = .25, *ns*). This is not surprising considering that the correlations were based on a relatively small sample (*N* = 25 doublets per each form). Although non-significant, these results seem to indicate that higher RGL is associated with higher *p*-values (easier items). Future studies should replicate this analysis to examine the relationship between RGL and difficulty in SJTs. See Figure 3 for depictions of these correlations.

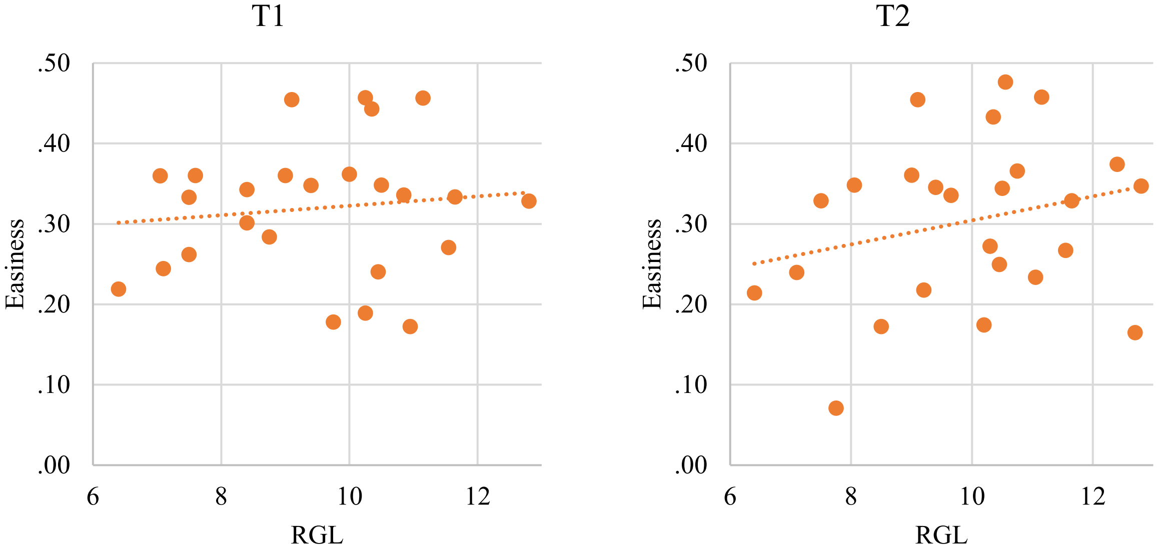


Figure 3. Doublet RGL and Easiness

4.0 SUBTEST-LEVEL RESULTS

This section covers the subtest-level analyses and results. The results reported in the text will be for the unstratified sample for the trichotomous scoring key for both Forms T1 and T2, except where noted. Results stratified by accession sources are provided in the appendices. As noted earlier, the subtest scores were calculated by averaging all 50 responses ('Overall' subtest), 25 ME responses ('ME' subtest), 25 LE responses ('LE' subtest), and the responses mapped to each of the six competencies.

4.1 Item-to-Competency Mapping

One of the objectives of the current research was to determine the extent to which the SJT measured the six identified competencies. As noted before, there are some reasons to believe the SJT may not adequately assess these competencies and that the psychometric properties of the competencies as subtests may be less than optimal. First, the SJT item development followed the CIT approach rather than the construct-driven approach (Barron, 2013). Therefore, the original item-to-competency mapping provided by the AFPC/DSYX was derived based on rational (not construct-based) considerations. IST performed an independent three-rater content analysis of the SJT items to map all items to the six competencies and to look for discrepancies with the original mapping. Indeed, seven discrepancies were found. Discussion with the AFPC/DSYX resulted in the adoption of the IST mapping. The subtest scores reported here are based on the IST item-to-competency mapping.

Second, the item-to-competency ratio was unbalanced both within and between the two forms. For example, Integrity was measured with five situations on Form T1 and with only two situations on Form T2 resulting in an unbalanced competency between forms. Additionally, Innovation was measured with two situations for Form T1 and Leading Others was measured with 7 situations for Form T1 resulting in unbalanced competencies within forms.

Finally, the distribution of the items within and between the forms was uneven. Sometimes the items loading onto the same competency were spread out (e.g., Leading Others in Form T1 included 1st and 24th item); other times they were clustered together (e.g., Innovation in Form T1 included items 11 and 16). Therefore, psychometric properties of the various subtests may be affected by order effects, length-fatigue, or length-related IER (Bowling et al., 2021). All of these factors contribute to the instability of the psychometric findings for the various competencies. Therefore, the statistics associated with these subtest scores should be interpreted with caution. The results for the other three subtest scores (Overall, ME, and LE) can be interpreted with greater confidence than the individual competency subtest results. See Appendix K for more detail.

4.2 Subtest Descriptives

For complete subtest-level descriptive statistics for the unstratified sample, see Table 12. For complete percentiles for the unstratified sample, see Table 13. In general, the scores on the subtests indicate the assessment is relatively easy with only moderate variance. Subtest scores were generally negatively skewed. The percentile distributions indicated that the SJT provided only low to moderate score differences among test-takers across different ability levels. For example, For Form T1, the mean Overall score for 1st percentile is 2.26 and for the 99th percentile 2.90. For stratified samples subtest-level descriptive statistics, see Appendix L.

Table 12. Unstratified Sample Subtest-Level Descriptives

SUBTEST	T1					T2				
	N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
Overall	35,751	2.66	.55	-2.73	12.49	32,548	2.67	.54	-2.51	9.61
Most Effective	35,749	2.63	.62	-1.80	3.06	32,544	2.62	.59	-1.81	4.01
Least Effective	35,752	2.70	.49	-3.66	21.91	32,553	2.71	.48	-3.20	15.20
Integrity	35,700	2.73	.54	-3.62	23.36	32,319	2.53	.61	-1.63	3.14
Leading Others	35,771	2.66	.53	-2.96	16.08	32,682	2.73	.51	-3.51	21.63
Decision Making	35,988	2.52	.59	-2.05	5.60	32,418	2.57	.50	-2.33	8.97
Communication Skills	35,445	2.66	.59	-2.45	7.29	32,378	2.66	.59	-2.54	8.02
Leading Innovation	35,987	2.76	.57	-2.71	8.06	32,854	2.80	.51	-2.44	4.85
Mentoring Others	35,651	2.69	.53	-2.11	5.60	32,749	2.70	.51	-2.06	4.81

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

Table 13. Unstratified Sample Subtest-Level Percentiles

SUBTEST							T1						
	99	95	90	80	70	60	50	40	30	20	10	5	1
Overall	2.90	2.84	2.81	2.77	2.74	2.70	2.68	2.65	2.62	2.56	2.50	2.42	2.26
Most Effective	2.92	2.88	2.84	2.76	2.72	2.68	2.64	2.60	2.56	2.48	2.40	2.32	2.16
Least Effective	2.96	2.88	2.84	2.80	2.76	2.76	2.72	2.68	2.64	2.60	2.52	2.44	2.28
Integrity	3.00	3.00	3.00	2.90	2.90	2.80	2.80	2.70	2.60	2.60	2.40	2.30	2.00
Leading Others	3.00	2.93	2.86	2.83	2.79	2.71	2.71	2.64	2.57	2.50	2.43	2.36	2.15
Decision Making	2.88	2.88	2.88	2.75	2.63	2.63	2.50	2.50	2.38	2.25	2.13	2.00	1.75
Communication Skills	3.00	3.00	3.00	3.00	2.83	2.83	2.67	2.67	2.50	2.33	2.25	2.00	1.75
Leading Innovation	3.00	3.00	3.00	3.00	3.00	3.00	3.00	2.75	2.50	2.50	2.25	2.00	1.75
Mentoring Others	3.00	3.00	3.00	2.88	2.88	2.75	2.75	2.63	2.63	2.50	2.38	2.25	2.00
SUBTEST							T2						
	99	95	90	80	70	60	50	40	30	20	10	5	1
Overall	2.88	2.84	2.82	2.78	2.74	2.72	2.68	2.66	2.62	2.56	2.48	2.40	2.22
Most Effective	2.92	2.84	2.80	2.76	2.72	2.68	2.64	2.60	2.56	2.48	2.40	2.29	2.12
Least Effective	2.92	2.88	2.88	2.84	2.80	2.76	2.72	2.72	2.68	2.60	2.52	2.44	2.22
Integrity	3.00	3.00	2.75	2.75	2.75	2.75	2.75	2.50	2.25	2.25	2.00	1.75	1.25
Leading Others	3.00	3.00	3.00	2.92	2.83	2.83	2.75	2.75	2.67	2.58	2.45	2.33	2.08
Decision Making	2.88	2.88	2.88	2.75	2.75	2.63	2.63	2.50	2.50	2.38	2.25	2.13	2.00
Communication Skills	3.00	2.92	2.92	2.83	2.75	2.73	2.67	2.58	2.56	2.50	2.33	2.25	2.00
Leading Innovation	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	2.50	2.00	2.00	1.50
Mentoring Others	3.00	2.92	2.92	2.83	2.83	2.75	2.75	2.67	2.67	2.58	2.42	2.33	2.08

4.3 Subtest Difficulty

Table 14 displays the subtest-level difficulty statistics for the unstratified sample. The SJT is a relatively easy assessment with mean *p*-values for the subtest scores ranging from .63 (T2: Integrity) to .86 (Forms T1 and T2: Innovation). For the stratified samples subtest-level difficulty statistics, see Appendix M.

Table 14. Unstratified Sample Subtest-Level Difficulty

SUBTEST	T1			T2		
	Min	Max	*M*	Min	Max	*M*
Overall	.44	.99	.78	.34	.99	.76
Most Effective	.45	.94	.76	.34	.97	.73
Least Effective	.44	.99	.80	.37	.99	.79
Integrity	.56	.99	.83	.34	.93	.63
Leading Others	.44	.99	.79	.55	.99	.82
Decision Making	.50	.96	.70	.49	.97	.69
Communication Skills	.48	.95	.78	.46	.97	.78
Leading Innovation	.81	.97	.86	.86	.87	.86
Mentoring Others	.52	.94	.75	.44	.94	.76

Note. Min = Minimum; Max = Maximum; *M* = Mean.

4.4 Subtest Discriminability

Table 15 displays the subtest-level discriminability statistics (i.e., ITCs) for the unstratified sample. The ITCs suggest that the SJT provides low discriminability and does not strongly differentiate between test-takers with low and high standing on the latent trait. Mean ITCs ranged from .11 (T1: Leading Others and Innovation) to .17 (T1: Mentoring Others; T2: LE and Integrity). Notice that some of the smallest ITCs were negative which suggests that the responses that make up that subtest do not hang together and may require deletion (T1: Response 47; T2: Response 7). Because the Innovation competency for Form T2 only contained two responses, discriminability was not calculated for this competency. For subtest-level discriminability statistics for the stratified samples, see Appendix M.

Table 15. Unstratified Sample Subtest-Level Discriminability

SUBTEST	T1			T2		
	Min	Max	*M*	Min	Max	*M*
Overall	-.04	.24	.14	-.02	.32	.16
Most Effective	-.04	.24	.13	-.02	.32	.16
Least Effective	.00	.24	.14	.07	.28	.17
Integrity	.09	.23	.14	.12	.21	.17
Leading Others	-.04	.20	.11	.05	.21	.13
Decision Making	.05	.24	.12	.08	.16	.12
Communication Skills	.10	.24	.16	-.03	.28	.14
Leading Innovation	.08	.12	.11	NA	NA	NA
Mentoring Others	.09	.22	.17	.08	.27	.15

Note. Min = Minimum; Max = Maximum; *M* = Mean. NA indicates that the parameters could not be estimated, because the Leading Innovation competency was measured by only 1 situation on T2.

4.5 Subtest Internal Consistency

Reliability at the item level (or internal consistency) was briefly mentioned in Section 3.4. This section concerns the reliability at the subtest level. Traditional methods for estimating SJT reliability like Cronbach's alpha may not be appropriate due to SJT's multidimensionality (Martin-

Raugh et al., 2018; Schäpers et al., 2020; Zinbarg, et al., 2005). Another reliability statistic known as McDonald's omega is usually recommended (Schäpers et al., 2020).

Total coefficient omega (ω) is an estimate of the general factor saturation of a test. It is comparable to Cronbach's alpha, but is generally considered a more accurate estimate of multidimensional test reliability (Zinbarg et al., 2005). Omega can change based on how the factors are estimated. The three methods include a minimum residual factor analysis, a principal axes factor analysis, and a maximum likelihood solution.

According to Zinbarg and colleagues, "A recommendation that should be heeded, regardless of the method chosen to estimate omega, is to always examine the pattern of the estimated general factor loadings prior to estimating omega. Such an examination constitutes an informal test of the assumption that there is a latent variable common to all of the scale's indicators that can be conducted even in the context of Exploratory Factor Analysis [EFA]. If the loadings were salient for only a relatively small subset of the indicators, this would suggest that there is no true general factor underlying the covariance matrix. Just such an informal assumption test would have afforded a great deal of protection against the possibility of misinterpreting the misleading omega estimates occasionally produced in the simulations reported here." (Zinbarg et al., 2005, p. 137).

Despite the recent research favoring omega over alpha for SJTs, many primary studies still report alpha. For that reason, we estimated both alpha *and* omega. Note that since omegas are grounded in the factor analytical framework, their estimation would only be possible for the subtests for which it is reasonable to assume that there is a general factor underlying the items. Due to the tentative nature of the item-to-competency mapping, omega is only estimated for the overall, ME, and LE subtests. Further, for stratified samples in which the solution could not converge, only alphas are presented.

Table 16 displays the Cronbach's alpha and McDonald's omega for the unstratified sample. As with most SJTs, Cronbach's alphas were below the acceptable levels (e.g., T1: $\alpha_{\text{Overall}} = .56$; T2: $\alpha_{\text{Overall}} = .64$; Lievens et al., 2008). The alphas between and within the parallel forms demonstrate considerable variance with the 'Overall' subtest having stronger alphas than ME, LE, or competency subtests. This is not surprising because alphas tend to increase as the number of items increases (Cortina, 1993). Note that alpha is not provided for the Form T2 Innovation competency due to a limited number of responses (i.e., 2). The 'Overall' subtest had the highest number of responses; therefore, its alphas were the highest. Omegas were slightly higher than alphas (T1: $\omega_{\text{Overall}} = .66$; T2: $\omega_{\text{Overall}} = .71$). For reliability information for the stratified samples, see Appendix N.

Table 16. Unstratified Sample Subtest-Level Internal Consistency

SUBTEST	T1		T2	
	α	*ω*	*α*	*ω*
Overall	.56	.66	.64	.71
Most Effective	.37	.47	.44	.53
Least Effective	.40	.51	.49	.58
Integrity	.31	NA	.31	NA
Leading Others	.19	NA	.30	NA
Decision Making	.33	NA	.28	NA
Communication Skills	.29	NA	.37	NA
Leading Innovation	.20	NA	NA	NA
Mentoring Others	.38	NA	.41	NA

Note. *α* = Cronbach's alpha; *ω* = McDonald's omega. NA indicates that the parameters could not be estimated due to factorial complexity of the subtests and due to the fact that the Leading Innovation competency was measured by only 1 situation on T2.

4.6 Subtest Subgroup Differences

As with the item-level analyses, the subgroup differences were assessed with Cohen's *d*, applying the same cutoffs. Table 17 displays the results of the subgroup differences for the unstratified sample. As can be seen, small to moderate Cohen's *d* at the response level compounded to produce higher Cohen's *d* at the subtest level, particularly for Black and Asian test-takers. The effect sizes were similar across the two forms, especially for the Overall, ME, and LE subtests; however, there were some notable discrepancies across forms. The most notable difference between T1 and T2 was observed for the Leading Others subtest, especially for the Black/White category (-.38). As previously mentioned, the competency-based statistics must be interpreted with caution because they were rationally derived based on the CIT rather than a construct-driven technique. Subtest subgroup differences for the stratified samples are available in Appendix O.

Table 17. Unstratified Sample Subtest-Level Subgroup Differences

SUBTEST	T1 Cohen's *d**					T2 Cohen's *d**				
	F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
Overall	.13	.68	.46	.25	.06	.03	.70	.46	.34	.01
Most Effective	.13	.63	.36	.23	.07	.09	.65	.35	.30	.02
Least Effective	.09	.50	.42	.18	.04	.05	.56	.45	.29	.04
Integrity	.08	.39	.23	.17	.03	.07	.27	.27	.17	.04
Leading Others	.08	.25	.12	.04	.08	.08	.63	.33	.24	.08
Decision Making	.20	.37	.36	.17	.06	.02	.31	.27	.13	.03
Communication Skills	.04	.41	.13	.14	.02	.05	.43	.14	.21	.04
Leading Innovation	.09	.29	.26	.08	.01	.09	.12	.14	.11	.00
Mentoring Others	.05	.45	.36	.18	.09	.04	.43	.37	.24	.05

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.
Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic.

4.7 Subtest Speededness

Speededness is an assessment of the extent to which the time allotted for a test drives the variance in the scores (Angoff, 1953). Presumably, when test-takers do not have enough time to finish a section, they are more likely to omit items and/or bubble in the same response option (e.g., “C”) for consecutive items in lieu of earnest attempts at those items. Since incorrect responses are not penalized on any of the AFOQT subtests, test-takers have no disincentive against using this strategy. Test-takers are allotted 35 minutes to complete the SJT, which equates to 84 seconds per item. To test for speededness, doublet longstring and response omissions were examined. While the SJT was not originally designed as a speeded test, it is worth seeing if there is evidence to suggest speededness.

Doublet longstring examines how often test-takers select the same course of action for a doublet (e.g., A for ME and A for LE). The percentage of doublet longstring increases steadily throughout the test for both forms (T1: starts with .04% and ends with 1.87%; T2: starts with .02% and ends with 2.13%).

Response omissions examine how often test-takers omit a response throughout the test. The review of the response omissions revealed that the percentage of the missing vales at the response level increases throughout the test (T1: starts with .07% and ends with 4.85%; T2: starts with .09% and ends with 5.96%). Figure 4 shows the percentage of response omissions for Forms T1 and T2.

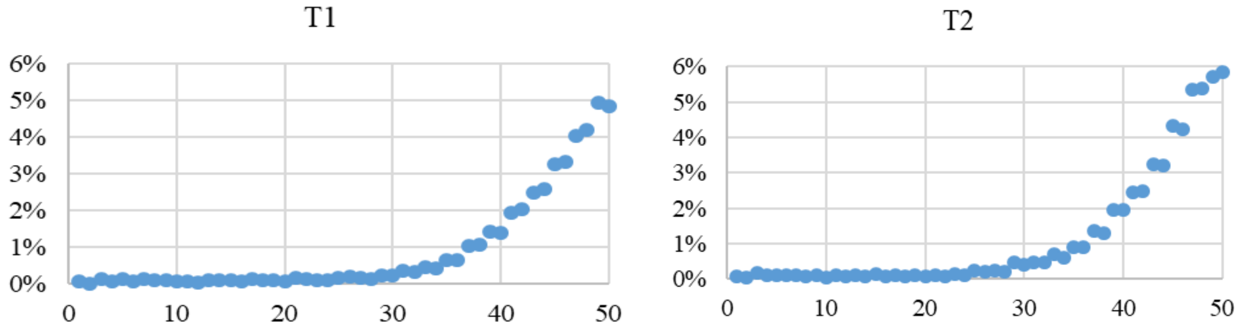


Figure 4. Response Omissions

Doublet omissions examine how often test-takers omit an entire doublet throughout the test. The percentage of the doublet omissions increases throughout the test (T1: starts with .05% and ends with 4.90%; T2: starts with .07% and ends with 5.79%). See Figure 5.

Together these results point to the possibility that some test-takers may be running out of time and either select responses that cannot possibly be correct (i.e., a doublet longstring) or omit responses altogether (i.e., response omission and doublet omission). Thus, we conclude that the SJT may be slightly speeded. However, longstrings, response omissions, and doublet omissions may also point to fatigue and low test-taking motivation, either independently or jointly. Test-takers may begin the subtest with the intention of responding with appropriate effort, but given the SJT is not part of any operational composites, motivation may wane toward the end of the subtest. There is not a way to disentangle what may be speededness from what may be other test-taking artifacts.

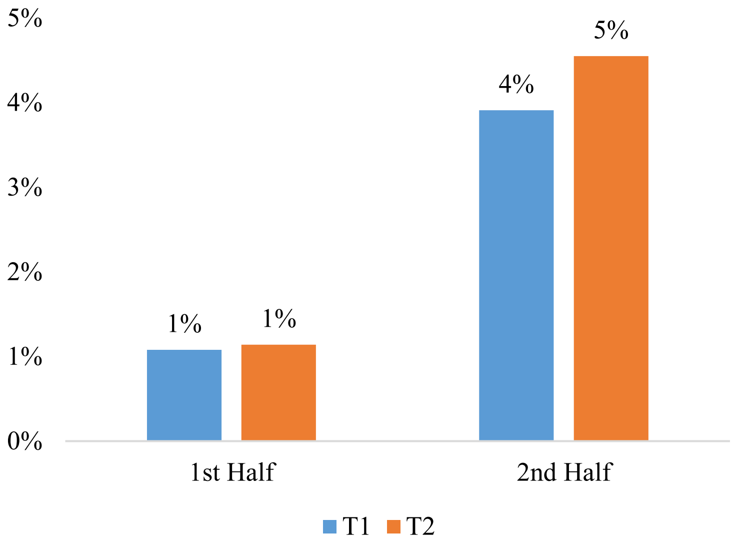


Figure 5. Percentage of at Least One Doublet Omission in Each Half of the Test

4.8 Subtest Correlations

Part of establishing the SJT construct validity (Cook et al., 2002) includes the examination of its discriminant and convergent validity. *Discriminant validity* refers to the lack of conceptual overlap between the two constructs (i.e., the constructs are considered unrelated; Goodwin & Leech, 2003). *Convergent validity* refers to a conceptual overlap between the two constructs (i.e., the constructs are considered related; Goodwin & Leech, 2003). To examine for potential discriminant and convergent validity, the SJT was correlated with the AFOQT cognitive subtests and the SDI-O. Because the SJT was developed as a strategy to reduce subgroup differences, ideally, correlations would be weaker with the cognitive subtests and stronger with the personality inventory. This pattern of relations would be ideal because cognitive ability is associated with higher subgroup differences and personality is associated with lower subgroup differences (Ployhart & Holtz, 2008). Presumably, if the SJT relates more strongly with personality than with cognitive ability, the subgroup differences for the SJT would also be lower.

4.8.1. SJT Subtest Inter-Correlations

Before examining the convergent and discriminant validity, the correlations among SJT subtests were computed. Correlations among the ME, LE, and Overall subtests as well as among the competencies were computed. The strongest correlations were those that included the ME, LE, and Overall subtests. These correlations are likely somewhat artificially inflated because each competency is also a part of the ME, LE, and Overall subtests. Of particular note is the modest correlation between ME and LE subtests (T1: $r = .45, p \leq .01$; T2: $r = .52, p \leq .01$). This continues to suggest that accuracy in evaluating effective and ineffective behaviors may be distinct skills, predicted by different constructs (Crook et al., 2011). The correlations among the competencies themselves were small to modest. See Table 18 for a complete SJT subtest scores correlation table.

Table 18. Unstratified Sample Subtest-Level Multivariate Correlations

SUBTEST	T1*								T2*							
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
Overall																
Most Effective	.88								.89							
Least Effective	.82	.45							.85	.52						
Integrity	.60	.61	.40						.47	.40	.43					
Leading Others	.60	.52	.49	.19					.68	.63	.55	.20				
Decision Making	.52	.42	.47	.14	.11				.52	.45	.46	.17	.20			
Communication Skills	.52	.50	.37	.19	.15	.12			.68	.69	.47	.19	.28	.20		
Leading Innovation	.37	.31	.33	.10	.09	.09	.10		.28	.18	.32	.06	.11	.08	.09	
Mentoring Others	.59	.43	.59	.24	.18	.16	.22	.15	.68	.53	.66	.20	.30	.21	.28	.16

*All correlations are significant at $p \leq .01$.

4.8.2 SJT Subtest Correlations with Cognitive Subtests

Next, to evaluate discriminant and convergent validity, correlations between the SJT and cognitive subtests were computed. The correlations ranged from small to moderate. The strongest correlations observed were those among the Overall, ME, and LE SJT subtests and the VA, WK, and RC cognitive subtests. This suggests substantial cognitive loading of the SJT. It is possible that the cognitive loading is a result of the instruction type currently in use for the SJT. The current SJT uses knowledge-based instructions rather than behavioral tendency instructions. Research suggests that the former instruction type tends to be correlated with cognitive ability more than with personality (Lievens et al., 2009; McDaniel et al., 2007). Table 19 provides the complete correlation table for the SJT subtest scores with the AFOQT cognitive subtests.

Table 19. Unstratified Sample Subtest-Level SJT and Cognitive Subtests Correlations

						T1			
SUBTEST	Overall	ME	LE	Integrity	Leading Others	Decision Making	Communication Skills	Leading Innovation	Mentoring Others
VA	.30	.28	.24	.15	.13	.20	.18	.14	.19
AR	.24	.22	.18	.13	.09	.17	.14	.09	.15
WK	.31	.28	.25	.14	.14	.21	.18	.14	.21
MK	.15	.15	.10	.10	.05	.12	.09	.06	.09
RC	.35	.32	.28	.18	.15	.24	.20	.16	.24
PS	.16	.16	.11	.08	.07	.13	.09	.05	.10
TR	.24	.22	.19	.16	.08	.12	.16	.10	.16
IC	.19	.18	.14	.12	.07	.15	.11	.05	.11
BC	.19	.18	.14	.11	.06	.13	.12	.07	.12
AI	.20	.18	.15	.12	.07	.15	.12	.05	.13
						T2			
SUBTEST	Overall	ME	LE	Integrity	Leading Others	Decision Making	Communication Skills	Leading Innovation	Mentoring Others
VA	.35	.34	.28	.16	.30	.17	.21	.06	.24
AR	.25	.25	.19	.10	.22	.13	.15	.04	.16
WK	.34	.33	.26	.15	.31	.14	.20	.06	.24
MK	.19	.19	.13	.07	.20	.09	.11	.02	.10
RC	.41	.38	.33	.18	.33	.20	.24	.09	.29
PS	.18	.19	.11	.05	.20	.06	.11	$.01^{\dagger}$	.11
TR	.27	.23	.23	.15	.18	.14	.18	.05	.17
IC	.19	.19	.13	.07	.18	.07	.13	.01*	.12
BC	.22	.21	.16	.10	.18	.10	.15	.02*	.14
AI	.19	.20	.13	.07	.17	.07	.13	.02*	.13

*All T1 correlations are significant at $p \leq .01$. All T2 correlations are significant at $p \leq .01$, except for Innovation-PS (*ns*), Innovation-IC ($p \leq .05$), Innovation-BC ($p \leq .05$), and Innovation-AI ($p \leq .05$).

4.8.3. SJT Subtest Correlations with SDI-O

The six SJT competencies should theoretically relate to at least some of the personality facets measured by the SDI-O. For example, it is conceivable that the SJT competency of Leading Others should relate to the Influence Tactics facet of the SDI-O. However, no formal hypotheses were made about specific relations between SJT subtest scores and SDI-O facets. All analyses were exploratory in nature. Furthermore, due to a data quality issue with four of the SDI-O facets, analyses for these facets should be interpreted with extreme caution.

Correlations between SJT subtest scores and SDI-O facets, SDI-O domains, Officer Suitability Measure (OSM) Facets, and the OSM score were small in magnitude (maximum absolute $r = .13$) with many being trivial (absolute r less than or equal to .05). Given the knowledge-based instruction type and its associated cognitive loading, it is not particularly surprising that the SJT does not meaningfully relate to many of the personality facets measured by the SDI-O. For a complete discussion of SDI-O data quality issues and resultant courses of action including original and corrected correlation tables, see Appendix P.

4.9 Retesting

The retesting policy for the AFOQT allows up to three attempts. As such, there was sufficient data to examine test-retest reliability and alternate forms reliability. This also allowed for comparisons between people who retest and those who do not retest, and comparisons between those who retest once and those who retest multiple times.

4.9.1. Test-Retest Reliability

As previously discussed, traditional internal consistency reliability statistics (i.e., Cronbach's alpha) may not be suitable for SJTs (Campion et al., 2014; Lievens et al., 2008). *Test-retest reliability* is suggested as a more appropriate statistic (Whetzel & McDaniel, 2009). Test-retest reliability is simply the correlation between individuals' scores on the first attempt and their scores on the subsequent attempt (computed from test-takers who have taken the same test form multiple times). Test-retest reliability was examined using data from test-takers who took the same form of the AFOQT twice. Test-retest reliability is rarely reported making interpretation of results difficult. Furthermore, the interval between the AFOQT administrations was much longer than what is typically reported in the literature. Campion and colleagues (2014) reported a mean test-retest reliability of .61 with a standard deviation of .26. Therefore, the mean test-retest reliability for the AFOQT SJT of .56 appears reasonable. Table 20 provides test-retest reliability for the unstratified sample. See Appendix Q for stratified samples test-retest reliability.

Table 20. Unstratified Sample Subtest-Level Test-Retest Reliability

SUBTEST	T1* r_{T1T1}	T2* r_{T2T2}
Overall	.64	.65
Most Effective	.64	.62
Least Effective	.59	.59
Integrity	.57	.53
Leading Others	.47	.59
Decision Making	.53	.38
Communication Skills	.53	.52
Leading Innovation	.52	.41
Mentoring Others	.62	.57

*All correlations are significant at $p \leq .01$.
Note. r = Correlation. N_{T1} = 1,422; N_{T2} = 1,128.

4.9.2. Alternate Forms Reliability

In addition to test-retest reliability, *alternate forms reliability* was computed. It serves as a way to determine how scores on one form correlate with the scores on the other form (Schmitt & Chan, 1998). It can also serve as a way to determine the equivalency of the parallel test forms (Murphy & DeShon, 2000).

Alternate forms reliability was assessed in three ways. First, between-subjects mean differences were examined between Forms T1 and T2 using independent samples t-tests. Second, between-subjects mean differences for common items were examined between Forms T1 and T2 using independent samples t-tests. As previously discussed, items did not appear in the same order on both assessments; therefore, results should be interpreted with caution due to the potential for order effects. Third, within-subject mean differences were examined for the test-takers who took Form T1 and then T2 or who took T2 and then T1 using paired samples t-tests. For this analysis, the order of assessment (i.e., T1 then T2 vs. T2 then T1) was balanced so the sample sizes were equal.

For analyses with unequal variance as indicated by Levene's test, Welch's *t*-test was used instead of the traditional Student's *t*-test (Levene, 1961).

The results of the independent samples *t*-tests with all items included, and with only common item included, revealed small to moderate effect sizes. The largest of these effect sizes was observed for Integrity (.69). As previously discussed, the results for the competencies should be interpreted with caution. In the case of alternate forms reliability, the imbalance between the number of items used to measure the competencies on Forms T1 and T2 is of particular concern. The effect sizes for the Overall/ME/LE subtests were small, suggesting that the two forms are parallel. Note that the mean differences were significant at $p \leq .01$, which is attributable to the large sample sizes, and do not indicate practically meaningful inequality between the two forms. A similar pattern of results was observed for the within-subjects paired samples *t*-tests. Effect sizes were small, and absolute mean differences were trivial.

The correlations between the two forms were small to moderate, with the largest being observed for the Overall subtest (r = .50, $p \leq .01$). This suggests that the scores the test-takers obtain on Form T1 are correlated with the scores they obtain on T2. The correlations are modest, and do not approach conventional reliability cutoffs (Cortina, 1993). There are several reasons why correlations between the administration of one form and the administration of the second form to the same test-takers are modest. This could reflect (1) low test-taking motivation, (2) improved performance from one administration to the next, (3) significant forgetting from one administration to the next due to a substantial time interval between administrations (i.e., a minimum of 150 days), (4) the imbalance of the competencies between Forms T1 and T2, or (5) the non-cognitive nature of the SJT (i.e., no objectively correct response may lead test-takers to choose a different response on different administrations of the same item at greater rates than for cognitive measures).

Overall, these results indicate that the differences between forms, while statistically significant, were practically non-significant, suggesting that equating between the forms is not necessary. See Table 21 for alternate forms reliability statistics for the unstratified sample and Appendix R for alternate forms reliability statistics for the stratified samples.

Table 21. Unstratified Sample Subtest-Level Alternate Forms Reliability

SUBTEST	All Items Between Subjects*		Common Items Between Subjects*		All Items Within Subjects*		
	Effect Size	Absolute Mean Difference	Effect Size	Absolute Mean Difference	Effect Size	Absolute Mean Difference	r_{T1T2}
Overall	.00†	.00	.05	.01	.02†	.00	.50
Most Effective	.09	.02	.08	.02	.10	.02	.41
Least Effective	.10	.01	.00†	.00	.08	.01	.40
Integrity	.69	.21	.01†	.01	.54	.22	.17
Leading Others	.35	.07	.08	.02	.16	.04	.28
Decision Making	.23	.06	.00†	.00	.21	.07	.15
Communication Skills	.06	.01	.10	.03	.01†	.00	.25
Leading Innovation	.12	.04	.16	.07	.12	.06	.15
Mentoring Others	.06	.01	.02	.01	.07	.02	.35

*All unmarked effect sizes and correlations are significant at $p \leq .01$; $^{\dagger}p$ not significant.
Note. r = Correlation; Effect Size = Cohen's *d*. All Items Between Subjects N = 65,864; Common Items Between Subjects N = 65,864; All Items Within Subjects N = 9,932.

4.9.3. Retester Analysis

There were several questions to answer to determine if there is significant impact of retesting. These inquiries include: Is there a difference between those test-takers who chose to retest (referred to as 'Retesters') versus those who chose not to retest (referred to as 'Non-Retesters')? Do test-takers improve their SJT score when they do retest? How much do the test-takers' scores improve with each subsequent retest? To answer these questions, paired samples *t*-tests were performed.

Note that these analyses differ from the test-retest and alternate forms analyses described before. *Retesting* of T1/T1 and of T2/T2 evaluates the hypothesis that people get better between their earlier and later attempts. *Test-retest* evaluates a hypothesis that the test takers' score on their first attempt on T1/T2 is closely correlated with their second attempt on T1/T2. *Alternate forms between-subjects analyses* evaluates the hypothesis that test-taker scores on Forms T1 and T2 are not significantly different (no equating). *Alternate forms within-subjects analyses* evaluates the hypothesis that people who take T1 first and then T2, or who take T2 first and then T1 would show strong correlations between the two scores.

It is conceivable that people who scored lower on their first attempt would retest in order to improve their score either to meet eligibility requirements for commissioning, or to improve their chances at placement into a rated career field. A paired-samples t-test was run to test this hypothesis. Indeed, there were significant differences between Retesters and Non-Retesters. Retesters scored, on average, lower on their first testing attempt than those who do not choose to retest. This is consistent with results for the AFOQT subtests (Carretta & Ree, 1997). See Figure 6 and Table 22 for more details.

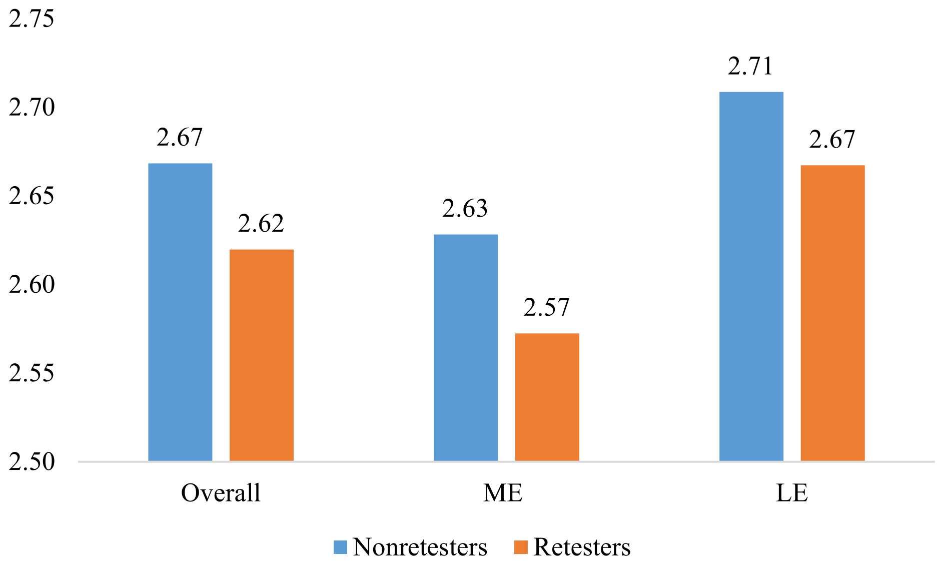


Figure 6. Mean Scores for Retesters and Non-Retesters

Table 22. Comparing Subtest-Level Retesters and Non-Retesters

SUBTEST	$M_{\text{Non-Retesters}}$	$SD_{\text{Non-Retesters}}$	$M_{\text{Retesters}}$	$SD_{\text{Retesters}}$	Effect Size*	Mean Difference
Overall	2.67	.13	2.62	.15	.36	.05
Most Effective	2.63	.17	2.57	.18	.33	.06
Least Effective	2.71	.14	2.67	.16	.29	.04
Integrity	2.63	.31	2.59	.34	.14	.04
Leading Others	2.71	.20	2.67	.21	.20	.04
Decision Making	2.55	.24	2.50	.26	.20	.05
Communication Skills	2.65	.26	2.58	.29	.25	.07
Leading Innovation	2.79	.36	2.74	.39	.12	.04
Mentoring Others	2.70	.22	2.65	.25	.25	.06

*All effect sizes are significant at $p \leq .01$.

Note. $M_{\text{Non-Retesters}}$ = Mean of Non-Retesters; $SD_{\text{Non-Retesters}}$ = Standard Deviation of Non-Retesters; $M_{\text{Retesters}}$ = Mean of Retesters; $SD_{\text{Retesters}}$ = Standard Deviation of Retesters; Effect Size = Cohen's *d*. $N_{\text{Non-Retesters}}$ = 63,749; $N_{\text{Retesters}}$ = 5,261.

It is also plausible that test-takers' scores would improve when they retest. However, unlike cognitive subtests, SJTs may be less amendable to score changes upon retesting. Literature suggests that coaching may help (Lievens et al., 2012). Coaching refers to test preparation tools that would teach people how to respond to SJT items most effectively (Lievens et al., 2012). However, as of the writing of this report, there are no USAF-wide coaching opportunities available for the test-takers. If such an opportunity were available, it would be optional. Additionally, the SJT is a low-stakes test by the virtue of not being included in the AFOQT operational composites. For these reasons, large mean score changes between the test attempts were not anticipated.

When test-takers retested on the same test form, their scores improved significantly, but the effect sizes of the increases in scores were small. This finding alleviates the concern for practice effects (i.e., that the test-takers scores would be higher on later attempts due to being exposed to the same test twice). Per the AFOQT policy, test-takers are instructed to take the opposite form each time they retest. With the maximum of three testing attempts, it is inevitable that the test-takers who retake the test twice will be exposed to the same test twice (e.g., T1 then T2 then T1 again). However, the minimum interval between test attempts is 150 days. Thus, the minimum interval between retesting on the same form should be 300 days. With almost a year between attempts on the same form, recall and practice effects should be minimal.

When retesting on the opposite forms, the test-takers' scores also improve significantly, but again the effect sizes of the increases in scores are small. This further confirms that the two forms are indeed parallel and do not require equating. See Figure 7 and Table 23 for more detail. See Appendix S for the stratified samples retesting effects.

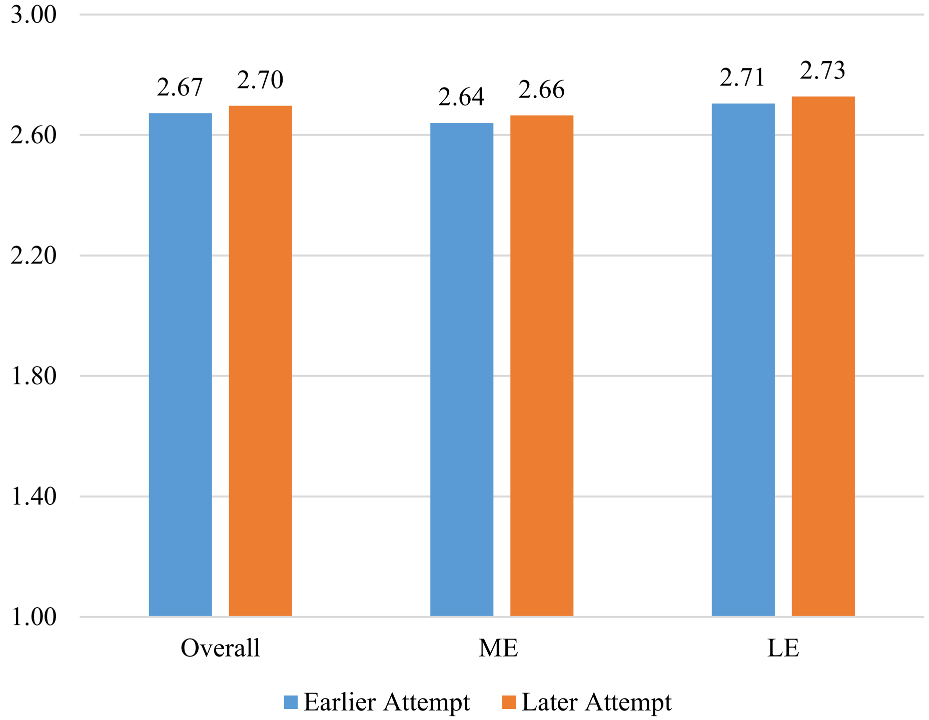


Figure 7. Mean Scores for Retesting on the Earlier and Later Attempt

Table 23. Unstratified Sample Retesting Effects on the Same and Parallel Forms

SUBTEST	T1-T1 (same form)								
	M_1	SD_1	M_2	SD_2	t	df	Effect Size	Mean Difference	r_{T1T1}
Overall	2.62	.16	2.64	.15	-4.39	710	.16**	.02	.64**
Most Effective	2.59	.19	2.60	.19	-2.68	710	.10**	.02	.64**
Least Effective	2.65	.17	2.68	.16	-4.87	710	.18**	.03	.59**
Integrity	2.69	.26	2.71	.25	-2.62	710	.10**	.02	.57**
Leading Others	2.65	.19	2.65	.20	.29	710	.01†	.00	.47**
Decision Making	2.46	.30	2.49	.28	-2.41	710	.09*	.03	.53**
Communication Skills	2.59	.34	2.64	.31	-4.26	710	.16**	.05	.53**
Leading Innovation	2.69	.37	2.71	.36	-.85	710	.03†	.01	.52**
Mentoring Others	2.63	.27	2.67	.26	-4.80	710	.18**	.04	.62**
SUBTEST	T2-T2 (same form)								
	M_1	SD_1	M_2	SD_2	t	df	Effect Size	Mean Difference	r_{T2T2}
Overall	2.61	.15	2.63	.16	-3.07	563	.13**	.02	.65**
Most Effective	2.55	.18	2.56	.18	-1.79	563	.08†	.01	.62**
Least Effective	2.67	.17	2.69	.17	-3.42	563	.14**	.02	.59**
Integrity	2.48	.41	2.51	.38	-1.69	560	.07†	.03	.53**
Leading Others	2.67	.24	2.69	.23	-1.73	563	.07†	.02	.59**
Decision Making	2.53	.22	2.53	.24	.33	563	.01**	.00	.38**
Communication Skills	2.58	.22	2.60	.23	-2.39	563	.10*	.02	.52**
Leading Innovation	2.76	.44	2.74	.47	.86	561	.04†	.02	.41**
Mentoring Others	2.65	.23	2.68	.21	-3.50	563	.15**	.03	.57**
SUBTEST	T1/T2 or T2/T1 (parallel forms)								
	M_1	SD_1	M_2	SD_2	t	df	Effect Size	Mean Difference	r_{T1T2}
Overall	2.62	.15	2.64	.15	-11.16	4965	.16**	.02	.51**
Most Effective	2.57	.18	2.59	.18	-8.42	4965	.12**	.02	.41**
Least Effective	2.67	.16	2.69	.16	-9.31	4965	.13**	.02	.40**
Integrity	2.57	.35	2.61	.32	-5.26	4950	.07**	.03	.03*
Leading Others	2.66	.21	2.67	.21	-2.03	4965	.03*	.01	.26**
Decision Making	2.50	.26	2.52	.26	-3.39	4965	.05**	.02	.13**
Communication Skills	2.58	.30	2.62	.29	-7.68	4965	.11**	.04	.25**
Leading Innovation	2.74	.39	2.72	.40	2.92	4953	.04**	.02	.14**
Mentoring Others	2.64	.25	2.68	.23	-10.16	4965	.14**	.04	.35**

*$p \le .05$; **$p \le .01$; $^{\dagger}p$ not significant.

Note. M_1 and SD_1 refer to the means and standard deviations for the earlier attempt. M_2 and SD_2 refer to the means and standard deviations of the later attempt. t = t-statistic; df = Degrees of Freedom; Effect Size = Cohen's d; r = Correlation. N_{T1T1} = 1,422; N_{T2T2} = 1128; $N_{T1T2orT2T1}$ = 9,932.

To assess how much test-takers improve when they retest, we examined the means of each attempt. Test-takers improved their SJT scores minimally on subsequent test-attempts. The magnitude of improvement seems to decrease on the third attempt. This could be evidence to suggest that more than one retest on the AFOQT provides very little increase in scores. Table 24 displays the means and standard deviations for each attempt for the Overall subtest only. Similar results were obtained for the rest of the SJT subtests.

Table 24. Overall Subtest Score Improvement over Multiple Attempts

Attempt	Statistic	Took Test Two Times	Took Test Three Times
1st Attempt	*M*	2.62	2.58
	SD	.15	.15
2nd Attempt	*M*	2.65	2.60
	SD	.15	.14
3rd Attempt	*M*	NA	2.61
	SD	NA	.16

Note. N for those who took the test two times 5,427. *N* for those who took test three times 564. NA indicates that the data were not available.

4.10 Test Security (Stability Analysis)

Test security is always a concern especially for enterprise-wide high-stakes employment tests such as the AFOQT. Although proactive steps have been taken to ensure test security (e.g., creation of alternate forms), it is prudent to continue scrutinizing the test regardless of test media (i.e., paper-and-pencil format or electronic format) for potential indicators of test security breach. Statistically, we test for breach of test security by examining whether or not there is a change in test parameters over time (similar to item drift).

Test security was examined by comparing the mean SJT scores for the test-takers who took the AFOQT between January 2015 and August 2016 and those who took the AFOQT between May 2018 and December 2019. Overall, some of the subtest score mean differences were statistically significant, however, due to large sample size, this is to be expected. Effect sizes and absolute mean differences were trivial with Cohen's *d* less than .20 and absolute mean differences less than .05 indicating no breach of test security. See Table 25 for the unstratified sample results and Appendix T for the results for the stratified samples. Note that stability analysis was not possible for the AECP stratified sample due to its small sample (*N* < 10).

Table 25. Unstratified Sample Subtest-Level Stability Analysis

SUBTEST	T1							
	M_{Early}	SD_{Early}	M_{Later}	SD_{Later}	t	df	Effect Size	Absolute Mean Difference
Overall	2.67	.13	2.66	.13	6.95	17998	.10	.01
Most Effective	2.64	.16	2.63	.17	4.77	17998	.07	.01
Least Effective	2.71	.14	2.69	.14	7.27	17998	.11	.02
Integrity	2.72	.22	2.73	.23	-.31	17998	.00†	.00
Leading Others	2.67	.18	2.67	.18	1.06	17998	.02†	.00
Decision Making	2.54	.26	2.50	.27	8.99	17997	.13	.04
Communication Skills	2.67	.30	2.64	.31	5.77	17997	.09	.03
Leading Innovation	2.78	.31	2.76	.33	4.11	17995	.06	.02
Mentoring Others	2.70	.24	2.68	.25	4.32	17998	.06	.02
SUBTEST	T2							
	M_{Early}	SD_{Early}	M_{Later}	SD_{Later}	t	df	Effect Size	Absolute Mean Difference
Overall	2.67	.13	2.66	.13	6.95	17998	.10	.01
Most Effective	2.64	.16	2.63	.17	4.77	17998	.07	.01
Least Effective	2.71	.14	2.69	.14	7.27	17998	.11	.02
Integrity	2.72	.22	2.73	.23	-.31	17998	.00†	.00
Leading Others	2.67	.18	2.67	.18	1.06	17998	.02†	.00
Decision Making	2.54	.26	2.50	.27	8.99	17997	.13	.04
Communication Skills	2.67	.30	2.64	.31	5.77	17997	.09	.03
Leading Innovation	2.78	.31	2.76	.33	4.11	17995	.06	.02
Mentoring Others	2.70	.24	2.68	.25	4.32	17998	.06	.02

*All unmarked effect seizes are significant at $p \leq .01$; †p not significant.

Note. M_{Early} = Mean of Earlier Group; SD_{Early} = Standard Deviation of Earlier Group; M_{Later} = Mean of Later Group; SD_{Later} = Standard Deviation of Later Group; t = t-statistic; df = Degrees of Freedom; Effect Size = Cohen's d. N_{T1} = 9,000; N_{T2} = 8,000.

5.0 QUALITATIVE REVIEW

The psychometric properties presented thus far apply to the current SJT design. Because SJTs are a measurement method and not a construct (Schmitt & Chan, 2006), there are many considerations that one should take into account when designing an SJT. There is no doubt that careful thought was put into the design of the current SJT and it shows potential for inclusion in the AFOQT composites. However, substantial room for improvement is evidenced by the item-to-competency imbalance both within and between forms, the fact that the SJT was based on CIT and not on a construct-driven technique, and the rather high correlations with the cognitive subtests.

Modifying SJT design features can influence its psychometric properties (e.g., Campion et al., 2014; Sullivan et al., 2019). Therefore, we performed a qualitative review of the current SJT attributes based on the Campion et al. (2014) taxonomy to explore the features of the current SJT and what changes are possible to improve its psychometric properties.

The situations and response options for the current SJT were generated using CIT. The most recent research recommends following the construct-driven approach because it results in a clearer factor structure and therefore better psychometric integrity and interpretability (Tiffin et al., 2020).
The current SJT uses an expert-based key development method and items are currently scored using a trichotomous approach. Although there are multiple SJT scoring methods described in the literature, no consensus has been reached about which one is best (Bergman et al., 2006; De Leng et al., 2019; McDaniel et al., 2011). Given the lack of a clear superior keying method, the choice of the scoring method should be continually evaluated and guided by the most recent research, but also by practical considerations. In the absence of strong psychometrically sound criteria, empirical, theoretical, and hybridized scoring methods may prove impractical. The other two scoring methods: expert-based and consensus (discussed below) may be more appropriate for the SJT in the present and near future.

The situation presentation for the current SJT is paper-and-pencil with a written stimulus and response media. The current SJT in its paper-and-pencil format was compared to a video-based SJT (Barron, 2013) and it was found that the latter does not outperform the former. There are studies to show that various multimedia SJTs may reduce subgroup difference by reducing reliance on verbal ability (Lievens & Sackett, 2006).

The response format for the current SJT is dual multiple choice using ME and LE responses. Arthur et al. (2014) compared the three response formats – rate, rank, and most/least – and found that depending on the goals of the SJT, each of these methods may prove useful. *Rate* response formats ask the test-takers to rate each response option using a Likert-type scale (the response options are therefore locally independent and normative); *rank* response formats ask test-takers to rank order the response options in some fashion (the responses are locally dependent and ipsative); *most/least* response formats ask test-takers to select most effective or most likely and least effective or least likely response options (the responses are locally dependent and ipsative).

The context for the current SJT is military-heavy (i.e., job specific). According to the Implicit Trait Policy (ITP) theory, job-specific knowledge may be more closely related to cognitive ability whereas job-nonspecific knowledge maybe more closely related to personality (Motowidlo & Beier, 2010). Therefore, it is conceivable that the military-heavy context would require in-depth knowledge of the military culture to determine what would be considered effective and ineffective. Although face valid, this knowledge can be obtained on the job and therefore does not need to be tested for on the AFOQT. It may also unfairly discriminate against test-takers who do not have substantial military knowledge in ways that are not job-relevant (e.g. the OTS-CIV or ROTC populations). Conversely, de-militarized items are more likely to tap into personality rather than cognitive ability and appropriately discriminate among test-takers based on job-relevant dimensions.

The constructs assessed in the current SJT are heterogeneous. The original intent was to measure six competencies. It is possible that a smaller, more targeted subset of these or new competencies would allow for balanced item-to-competency ratios within and between forms while still allowing for sufficient construct coverage. These modifications may allow for improved psychometric properties including factor structure, reliability, and correlations with external criteria.

The original research design was to collect the predictor data from test-takers as applicants and then correlate these scores with criteria for the same test-takers once they have become incumbents (i.e., commissioned officers). This is a predictive criterion validation design (Schmitt & Chan, 1998). This design is less commonly researched due to the practical difficulties involved in data collection, and is often abandoned in favor of concurrent validation. However, predictive validity is the gold-standard in validation design due to less serious concerns about range restriction with a predictive design than with a concurrent design. Therefore, it is recommended that the predictive validation design continue to be employed.

6.0 DISCUSSION

This technical report describes the results of the item-level and the subtest-level psychometric evaluation of the AFOQT SJT. The SJT was included in the AFOQT following the job analytic study by Lentz et al. (2009a; 2009b) with the goals to (1) expand the competencies measured by the AFOQT, (2) incrementally improve criterion-related validity above and beyond the AFOQT cognitive subtests, and (3) meet the USAF D&I goals by reducing subgroup differences (Barron, 2013; Lentz et al., 2009a; 2009b). In its current form, the SJT serves as an experimental subtest and is not included in any of the operational composites, which means test-taking motivation may be low for it, which may affect data quality and the conclusions, which may be reached. The paragraphs below summarize our findings and issue recommendations for how to improve the SJT in future iterations of the AFOQT.

Consistent with the initial validation study by Barron (2013), the current SJT shows promise in terms of expanding the AFOQT officership competency coverage and reducing subgroup differences, however, there is room for improvement.

A review of the various scoring methods (dichotomous, trichotomous, and continuous) revealed that the trichotomous scoring method was acceptable. While none of the scoring methods were a clear frontrunner in terms of internal psychometrics and criterion-related correlations, the trichotomous scoring method was recommended because it is currently used for the SJT and barring conclusive evidence of a better scoring option, there is no evidence to change the scoring methods. Continuing to use the same scoring method makes the interpretation of the existing scores easier and more intuitive and comparison of scores across forms simpler.

At the item-level, the SJT appears relatively easy and does not differentiate well among test-takers with high and low latent trait standing. Omission and longstring rates indicated that the test may be slightly speeded. Subgroup mean score differences were small to moderate, which is an improvement over the AFOQT cognitive subtests. Item drift analysis showed no evidence for test security breach. RGL analysis suggests that the SJT's reading level varies between and within forms with a range between 9th and 10th grade. Analyses involving factor structure revealed a messy and inconclusive factor structure which suggests it is unclear what the SJT is measuring or intended to measure.

Subtest-level results indicate the SJT is easy, provides low to moderate discriminability, and shows low to moderate internal consistency. The SJT is correlated with the AFOQT cognitive subtests at a higher rate than with the SDI-O. The SJT has acceptable test-retest and alternate forms reliability, with no equating necessary. Finally, test-takers whose first-attempt scores were lower tend to retest at higher rates than test-takers whose first-attempt scores were higher. This is likely a consequence of test-takers trying to improve their scores on the AFOQT cognitive subtests, since the SJT is experimental and does not affect the operational composites. When test-takers do retest, their scores improve, albeit minimally. Stability analysis suggests that there was no breach of test security.

Given that the SJT is a method and not a construct (Schmitt & Chan, 2006), modifying its design features can influence its psychometric properties. The current SJT has knowledge-based instruction type with a most/least effective response format. While this design can help improve the SJT's criterion-related validity, it may also increase subgroup differences and render the responses locally dependent. We recommend that the instruction type be changed to behavioral tendency which has been shown to reduce subgroup differences and the response format be changed to a rate format, whereby test-takers are asked to rate each response option separately, using a Likert-type scale. This will render the responses locally independent, will help reduce response latency, and may improve subgroup differences.

We also recommend that the SJT situations and response options be designed based on the construct-driven method rather than using the CIT. Items designed around a shorter more targeted competency list may improve the SJT factor structure and other psychometric properties (e.g. reliability). Further consideration should be given to competencies that are required upon entry compared to those that could be developed on the job and whether identified competencies truly distinguish high performers from others. The militarized context of the SJT may be an impediment to proper measurement and ability to reduce adverse impact.

Finally, to reduce potential effects of speededness for the current SJT, it is recommended that the time limit be extended to allow for a comfortable test completion. Alternatively, if a shorter version of SJT is considered, a shorter time limit may be possible. A shorter SJT may enable a shorter testing time while maintaining measurement capacity because of the change to the rate format, which allows researchers to receive the same number of measurements from a smaller number of items. As the AFOQT evolves, consideration may be given to multimedia presentation of the SJT, but in its current paper-and-pencil format, this is not possible. All of these modifications support the goal of increasing incremental validity while reducing subgroup differences.

Based on the aforementioned findings and the goals of continually improving performance prediction while ameliorating mean score subgroup differences, we issue the following recommendations.

Instruction Type

First, we recommend using behavioral tendency instructions ('what would you do') instead of the knowledge-based instructions ('what should you do'). The subtle difference in the wording has been shown to produce profound outcome differences. Research suggests that behavioral tendency instructions have lower correlations with cognitive ability, higher correlations with personality, similar criterion-related validities in high-stakes testing environments, and lower subgroup differences than knowledge-based instructions (Sullivan et al., 2019; Lievens et al., 2009). However, by definition, knowledge-based items cannot be faked because the response is supposed to be an objective assessment of the most and least effective actions regardless of what test-takers

themselves would actually do. Conversely, behavioral tendency items can be faked because test-takers may respond with an action that they perceive to be more appropriate than the one they themselves would actually perform given the situation in order to get the item correct. Thus, faking is an obstacle that would need to be overcome with the use of behavioral tendency instructions.
It could be argued that the instructions for the current SJT do not include the words 'what should you do.' Instead, they read as: "For each situation, you must respond to two questions on your answer sheet. First, select which one of the five actions listed you judge the MOST EFFECTIVE action in response to the situation. Then, select which one of the five actions listed you judge the LEAST EFFECTIVE action in response to the situation." However, the mere absence of the words 'what should you do' does not make the measure any less cognitively loaded. In fact, asking the test-takers to "Select the single MOST EFFECTIVE action (A-E) in response to the situation" and "Select the single LEAST EFFECTIVE action (A-E) in response to the situation" pushes people to consider the best and worst responses options, rather than the response option they personally would take. See Table 26 for an example of a knowledge-based and a behavioral tendency item.

Table 26. SJT Instruction Type Examples

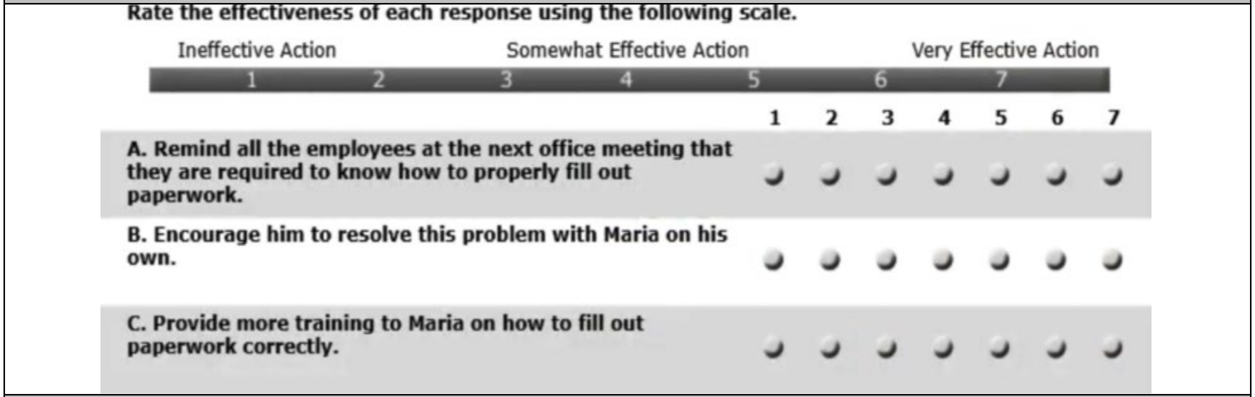
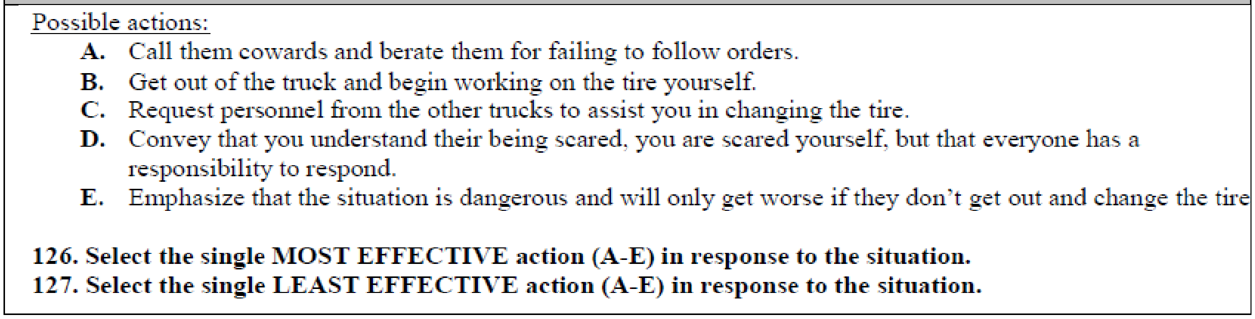
Knowledge-Based Item	Behavioral Tendency Item
You are a squad leader on a field exercise, and your squad is ready to bed down for the night. The tent has not been put up yet, and nobody in the squad wants to put up the tent. They all know that it would be the best place to sleep since it may rain, but they are tired and just want to go to bed. **What *should* you do?** A. Tell them that the first four men to volunteer to put up the tent will get light duty tomorrow. B. Make the squad sleep without tents. C. Tell them that they will all work together and put up the tent.	After you have served 2 years as manager of the sales team, the director of your company appoints a new deputy manager. Although you have been able to work together, your impressions of her are negative - you find her arrogant and disloyal. The director has now considered sending her on a course that would create an opportunity for her relocation to a different position within the company. However, it would also speed up her promotion. **What *would* you do?** A. You approve her participation in the course. B. You contact your director immediately and ask that she be relocated to a different position, more suited to her capabilities. C. You veto her participation in the course and discuss it with her.

Note. The example of a knowledge-based item was retrieved from National Research Council (2015). The example of a behavioral tendency item was retrieved from Jobtestprep.

Response Format

Second, we recommend the rate response format because it generally shows lower cognitive load, lower subgroup differences, and high internal consistency (Arthur et al., 2014; Ployhart & Ehrhart, 2002; Waugh & Russell, 2005). It is highly susceptible to faking and has low alternate forms reliability, which are limitations, which would have to be overcome. The rank response format generally shows the highest cognitive load, low internal consistency, moderate to high subgroup differences, poor applicant reactions, but modest susceptibility to faking (Arthur et al., 2014; Ployhart & Ehrhart, 2002; Waugh & Russell, 2005). The Most/Least format represents a middle of the road response format with respect to cognitive load and subgroup differences. It also has modest applicant reactions and the least susceptibility to faking (Arthur et al., 2014; Ployhart & Ehrhart, 2002; Waugh & Russell, 2005). The rate format best supports the objectives of reducing subgroup differences while increasing incremental prediction. See Table 27 for an example of each response format.

Table 27. SJT Response Format Examples

Rate
 Rate the effectiveness of each response using the following scale. Ineffective Action · Somewhat Effective Action · Very Effective Action 1 2 3 4 5 6 7 1 2 3 4 5 6 7 A. Remind all the employees at the next office meeting that they are required to know how to properly fill out paperwork. ○ ○ ○ ○ ○ ○ ○ B. Encourage him to resolve this problem with Maria on his own. ○ ○ ○ ○ ○ ○ ○ C. Provide more training to Maria on how to fill out paperwork correctly. ○ ○ ○ ○ ○ ○ ○
Rank
Please rank these potential actions in order of effectiveness with 1 being least effective and 5 being most effective. A. Since this course is likely to result in the relocation of the deputy manager, you approve her participation in the course. B. You contact your director immediately and ask that she be relocated to a different position, more suited to her capabilities. C. You veto her participation in the course and discuss it with her. You express your concerns and you try to work out your differences. You update your director.
Most/Least
 Possible actions: **A.** Call them cowards and berate them for failing to follow orders. **B.** Get out of the truck and begin working on the tire yourself. **C.** Request personnel from the other trucks to assist you in changing the tire. **D.** Convey that you understand their being scared, you are scared yourself, but that everyone has a responsibility to respond. **E.** Emphasize that the situation is dangerous and will only get worse if they don't get out and change the tire **126. Select the single MOST EFFECTIVE action (A-E) in response to the situation.** **127. Select the single LEAST EFFECTIVE action (A-E) in response to the situation.**

Scoring Method

Third, we recommend continuing to explore the viable scoring methods. As previously discussed, there are several options available. Expert-based keys are currently in use. However, the SME-based scores also produce larger sex subgroup differences and correlate more strongly with cognitive ability than do the novice/expert comparison scores (Bergman et al., 2006; De Leng et al., 2019; McDaniel et al., 2011). This may be due to the composition of the SME pool (e.g., largely male, extensive job-specific knowledge). These concerns may be mitigated by careful selection of SMEs. Empirical scoring methods were explored, but unfortunately did not produce interpretable results. A more recent scoring method, consensus scoring, may also be viable to provide adequate incremental validity while minimizing subgroup differences (Weng et al., 2018). These scoring methods should be explored in future iterations of the SJT.

Context

Fourth, as previously discussed, the current SJT context is military-heavy. We recommend de-militarizing the future SJT items. The military-heavy context was the result of generating critical incidents from military SMEs. Although face valid and potentially appropriate, we do think that the militarized context tends to tap into job-specific knowledge and in turn makes the SJT more cognitively-loaded, which may ultimately disadvantage certain groups. Extrapolating from ITP theory, it is likely that a more general, de-militarized item context may improve subgroup differences by the virtue of relying on fundamental socialization (i.e., job-non-specific knowledge).

Item-to-Competency Balance

Fifth, we recommend that the future SJT (1) achieves a balanced item-to-competency ratio both within and between the alternate forms and (2) either evenly or randomly (for each test-taker for each test administration) distributes the items throughout the test. Random distribution relies on electronic administration rather than paper-and-pencil. This balancing of competencies will certainly improve psychometric analysis of items and may improve psychometric properties.

Competencies

Sixth, we recommend reduction of the number of competencies that the SJT aims to measure. Even though SJTs are often multidimensional and may allow for the measurement for more than one construct, including six constructs clearly rendered the interpretation of some statistics (i.e., PCA) uninterpretable. Therefore, we recommend that a set of no more than four competencies be included in the SJT design and that the item and situation generation be construct-driven (Tiffin et al., 2020).

Speededness

Our final recommendation is to reduce the number of items overall to eliminate possible speededness. There is no indication that the SJT was ever intended to be a speeded subtest (Barron et al., 2013). With fewer competencies to measure and five responses per item resulting from the rate response format, we will be able to obtain more data with fewer items. Additionally, reducing the cognitive load by changing the instruction format and the response format will likely reduce the amount of time test-takers spend on each response. These should work together to address any speededness present in the SJT.

7.0 CONCLUSION

Although a great foundation was laid by including the SJT in the AFOQT as a way to meet the objectives of reduced subgroup differences, incremental validity, and competency coverage, the assessment as currently constructed is not without its shortcomings. Based on the present findings, a revision of the SJT is recommended. It is recommended that the next iteration of the SJT be based on the construct-driven method with behavioral tendency instruction type and rate response format. The new SJT should have a balanced item-to-competency ratio both between and within forms, and should focus on a shortened list of three or four competencies critical to officer performance and required upon entry into commissioned officership.

8.0 REFERENCES

Allison, P. D. (2001). *Missing data*. Sage publications.

Angoff, W. H. (1953). The reliability and effective test length. *Psychometrika*, *18*(1), 1-14.

Arthur, Jr., W., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology*, *99*(3), 535-545.

Barron, L. G. (2013). *Air Force Officer Qualifying Test (AFOQT) Situational Judgment Test (SJT) development*, AFCAPS-TR-2013-0005. Randolph AFB, TX: Air Force Personnel Center, Strategic Research and Analysis Branch.

Barron, L. G., Rose, M. R., Aguilar, I. D., & Carretta, T. R. (in press).Development of a situational judgment test to supplement current US Air Force measures of officership. *Military Psychology*.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*(3), 223-235.

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *24*(4), 718-738.

Campion, M. C., Ployhart, R. E., & MacKenzie, Jr, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance*, *27*(4), 283-310.

Carretta, T. R., & Ree, M. J. (1997). *The best retest is the average: Findings and implications*, AL/HR-TP-1996-0021. Brooks AFB, TX: Armstrong Laboratory Human Resources Directorate, Manpower and Personnel Research Division.

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). 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.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104.

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

Crook, A. E., Beier, M. E., Cox, C. B., Kell, H. J., Hanks, A. R., & Motowidlo, S. J. (2011). Measuring relationships between personality, knowledge, and performance using single-response situational judgment tests. International Journal of Selection and Assessment, 19(4), 363-373.

De Leng, W. E., Stegers-Jager, K. M., Born, M. P., & Themmen, A. P. N. (2019). Faking on a situational judgment test in a medical school selection setting: Effect of different scoring methods. *International Journal of Selection and Assessment*, *27*(3), 235-248.

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*, 1-18.

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

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484-501.

Goodwin, L. D., & Leech, N. L. (2003). The Meaning of Validity in the New Standards for Educational and Psychological Testing: Implications for Measurement Courses. *Measurement and evaluation in Counseling and Development*.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828-845.

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

Jobtestprep (2021). Retrieved on June 22, 2021 from Free Situational Judgment Questions and Answers jobtestprep.co.uk

Kantrowitz, T., Kingry, D., Engelsted, J., Travinin, G., Lovering, E., & Gould, M. (under review). *Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: Validity and Subgroup Differences for Current and Alternative Composites,* AFRL-RH-WP-TR-2021-xxxx. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Kline, T. J. (2005). Classical test theory: Assumptions, equations, limitations, and item analyses. *Psychological testing: A practical approach to design and evaluation*, *91*.

Kutner, M., Greenberg, E., Jin, Y. Boyle, B., Hsu, Y.-C., & Dunleavy, E. (2007, April). *Literacy in everyday life: Results from the 2003 national assessment of adult literacy.* (NCES 2007-480). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics.

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

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

Levene, H. (1961). Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling, 2*79-292.

Lievens, F., Buyse, T., Sackett, P. R., & Connelly, B. S. (2012). The effects of coaching on situational judgment tests in high-stakes selection. *International Journal of Selection and Assessment*, *20*(3), 272-282.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426-441.

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91*(5), 1181–1188. https://doi.org/10.1037/0021-9010.91.5.1181

Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology*, *94*(4), 1095-1101.

Martin-Raugh, M. P., Anguiano-Carrsaco, C., Jackson, T., Brenneman, M. W., Carney, L., Barnwell, P., & Kochert, J. (2018). Effects of Situational Judgment Test Format on Reliability and Validity. *International Journal of Testing*, *18*(2), 135-154.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb III, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology*, *60*(1), 63-91.

McDaniel, M. A., List, S. K., & Kepes, S. (2016). The "hot mess" of situational judgment test construct validity and other issues. *Industrial and Organizational Psychology*, *9*(1), 47-51.

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, *96*(2), 327-336.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, *17*(3), 437.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*(6), 640-647.

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology*, *95*(2), 321-333.

Murphy, K. R., & DeShon, R. (2000). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology*, *53*, 913-924.

National Research Council. (2015). *Measuring human capabilities: An agenda for basic research on the assessment of individual and group performance potential for military accession*. National Academies Press.

Nguyen, McDaniel, & Whetzel, 2005 judgment test performance: A meta-analysis. Paper presented at the 20th annual conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA

Ployhart, R. E., & Ehrhart, M. G. (2002). Modeling the practical effects of applicant reactions: subgroup differences in test-taking motivation, test performance, and selection rates. *International Journal of Selection and Assessment*, *10*(4), 258-270.

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

Ployhart, R. E., & MacKenzie, Jr., W. I. (2011). Situational judgment tests: A critical review and agenda for the future. *APA handbook of industrial and organizational psychology: Selecting and developing members for the organization* (Vol. 2, pp. 185-204) Washington, DC: American Psychological Association.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74-104.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147-177.

Schäpers, P., Mussel, P., Lievens, F., König, C. J., Freudenstein, J. P., & Krumm, S. (2020). The role of Situations in Situational Judgment Tests: Effects on construct saturation, predictive validity, and applicant perceptions. *Journal of Applied Psychology*, *105*(8), 800-818.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological assessment*, *8*(4), 350.

Schmitt, N., & Chan, D. (1998). *Personnel Selection: A theoretical approach*. Sage Publications, Inc.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct. *Situational judgment tests: Theory, measurement, and application*, 135-155

Sullivan, T. S., Whetzel, D. L., & McCloy, R. A. (2019). *Literature Summary: Best Practices in Situational Judgment Test (SJT) Development*. Human Resources Research Organization Alexandria United States.

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. https://doi.org/10.1111/j.1745-3984.1989.tb00326.x

Tiffin, P. A., Paton, L. W., O'Mara, D., MacCann, C., Lang, J. W., & Lievens, F. (2020). Situational judgement tests for selection: Traditional vs construct-driven approaches. *Medical education*, *54*(2), 105-115.

Walsh, J. L., Brady, M. F., Woolley, M. R., & Carretta, T. R. (under review, a). Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: Item-Level Analyses, AFRL-RH-WP-TR-2021-xxxx. Wright-Patterson AFB, OH: 711 Human PerformanceWing, Airman Biosciences Division, Performance Optimization Branch.

Walsh, J. L., Woolley, M. R., Brady, M. F., & Carretta, T. R. (under review). *Air Force Officer Qualifying Test (AFOQT) Form T Evaluation: Subtest-Level Analyses,* AFRL-RH-WP-TR-2021-xxxx. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Waugh, G. W., & Russell, T. L. (2005). Criterion situational judgment test (CSJT). *Development of experimental Army enlisted personnel selection and classification tests and performance criteria*.

Weng, Q. D., Yang, H., Lievens, F., & McDaniel, M. A. (2018). Optimizing the validity of situational judgment tests: The importance of scoring methods. *Journal of Vocational Behavior*, *104*, 199-209.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*(3), 188–202. https://doi.org/10.1016/j.hrmr.2009.03.007

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*(3), 291-309.

Woolley, M. R., Walsh, J. L., Mann, K. J., Wilson, R. T., & Carretta, T. R. (in progress). *Self-Description Inventory - Officer (SDI-O): Item-, facet-, & domain-level analyses,* AFRL-RH-WP-TR-2022-xxxx. Wright-Patterson AFB, OH: 711 Human Performance Wing, Airman Biosciences Division, Performance Optimization Branch.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133.

APPENDIX A: Competency Definitions

Competency	Definition
Displaying Integrity, Ethical Behavior, and Professionalism	Displaying uncompromising commitment to the Air Force core values; always maintaining ethical principles and telling the truth, regardless of consequences; accepting responsibility for own and subordinates' actions; maintaining sharp military appearance and physical health/fitness; supporting Air Force mission and goals; having a thorough understanding of military regulations and initiatives and carrying them out in accordance with Air Force standards; following policies, regulations, and orders, and defending them to others; understanding the chain of command, and accepting and respecting the decisions of superiors; displaying appropriate courtesies to others; understanding how policies and actions fit into the overall mission scheme.
Decision-Making and Managing Resources	Managing resources efficiently and effectively; ensuring deadlines are met through planning, and effectively utilizing resources; gathering information, identifying risks and goals, and assessing available resources to complete projects on time and within budget; prioritizing tasks; sorting through large quantities of information efficiently; focusing on multiple tasks and requirements; making sound decisions; appropriately considering relevant sides of an issue; remaining focused and decisive in stressful situations.
Mentoring Others	Providing guidance to subordinates and others; assessing strengths and weaknesses in personnel and providing them with honest and specific feedback; designing opportunities for subordinates to develop new skills, and assisting them in establishing career plans; providing subordinates with strategic vision and goals; sharing knowledge and experience with subordinates.
Pursuing Personal and Professional Development	Continuously improving professional skills, knowledge, and abilities through formal and informal training, professional military education, off-duty education, on-the-job training, etc.; ability to find purpose, personal growth in work; balancing professional development and training with job completion such that performance does not suffer; maintaining superior technical skills through training.
Leading Innovation	Being open to new ideas and new methods for accomplishing goals; adjusting to a rapidly changing environment and modifying goals and objectives based on emerging requirements; embracing innovation and looking for better methods/techniques to accomplish tasks; adapting to new and changing missions, tasks, and situations.
Leading Others	Effectively building and leading individual and team activities; persuading, inspiring, and motivating others, regardless of their relative positions in the hierarchy; creating a sense of enthusiasm and purpose in own team; demonstrating a positive attitude, and team spirit to inspire subordinates; effectively adopting different leadership styles as appropriate to individuals and settings.
Communication Skills	Practicing meaningful two-way communication (i.e., speaking and writing clearly, listening attentively and clarifying information); providing timely and relevant information up and down the chain of command; tailoring presentations to the level of the audience; expressing opinions when appropriate; expressing oneself in a manner that produces a productive and harmonious environment; ability to evaluate the importance of information being communicated.

APPENDIX B: Stratified Samples Demographics

Sample Characteristic	T1													
	OTS-CIV		OTS-AD		AECP		USAFA		ROTC		ANG		AFRES	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Sex														
Male	5,619	70	4,308	79	132	78	3,192	74	9,145	72	3,050	78	1,111	75
Female	2,397	30	1,112	20	38	22	1,108	26	3,603	28	863	22	367	25
Unknown	6	0	8	0	0	0	3	0	11	0	2	0	3	0
Race*														
AIAN	374	5	322	6	8	5	200	5	728	6	220	6	86	6
Asian	888	11	472	9	10	6	494	11	1,483	12	259	7	122	8
Black or AA	1,143	14	1,009	19	36	21	418	10	1,662	13	437	11	300	20
NHPI	165	2	164	3	7	4	141	3	332	3	82	2	27	2
White	6,141	77	3,982	73	119	70	3,634	84	9,688	76	3,243	83	1,033	70
Ethnicity														
Hispanic	1,143	14	803	15	33	19	491	11	2,092	16	416	11	230	16
Non-Hispanic	6,761	84	4,515	83	134	79	3,762	87	10,387	81	3,427	88	1,218	82
Unknown	118	1	110	2	3	2	50	1	280	2	72	2	33	2

Sample Characteristic	T2													
	OTS-CIV		OTS-AD		AECP		USAFA		ROTC		ANG		AFRES	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Sex														
Male	5,160	71	3,896	81	129	77	2,789	75	8,831	71	2,595	79	877	72
Female	2,114	29	939	19	38	23	944	25	3,550	29	691	21	337	28
Unknown	8	0	2	0	1	1	3	0	25	0	2	0	1	0
Race*														
AIAN	329	5	335	7	10	6	165	4	753	6	153	5	67	6
Asian	735	10	395	8	20	12	419	11	1,425	11	223	7	92	8
Black or AA	1,052	14	930	19	38	23	339	9	1,604	13	376	11	235	19
NHPI	157	2	150	3	3	2	103	3	311	3	67	2	33	3
White	5,511	76	3,473	72	121	72	3,158	85	9,426	76	2,703	82	858	71
Ethnicity														
Hispanic	1,043	14	707	15	22	13	398	11	2,058	17	334	10	180	15
Non-Hispanic	6,083	84	4,023	83	140	83	3,275	88	10,083	81	2,884	88	1,012	83
Unknown	156	2	107	2	6	4	63	2	265	2	70	2	23	2

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

Note. AIAN = American Indian/Alaska Native; AA = African American; NHPI = Native Hawaiian/Other Pacific Islander.

APPENDIX C. Statistics and Their Interpretation

CTT Statistic	Interpretation
Skewness	Skewness is a departure from normality in which the data is clustered to one side of the distribution or the other.
Kurtosis	Kurtosis is a departure from normality in which the data is more peaked and densely distributed or flatter and less densely distributed than a normal distribution.
Cronbach’s alpha (α)	A measure of internal consistency (reliability) of a psychological scale.
McDonald’s omega (ω)	Reliability estimate calculated from factor analytic models used to represent the association between an item on a scale and the scale’s purported construct.
Correlation (*r*)	A correlation is a relation between two (i.e., bivariate) or more (i.e., multivariate) variables. Positive correlations indicate that as one variable increases, the other variable increases in a predictable manner. Negative correlations indicate that as one variable increases, the other variable decreases in a predictable manner.
Item Difficulty (*p*-value)	In this context, a measure of difficulty. Represents the percentage of applicants who selected the correct answer. Higher *p*-values indicate easier items.
Item Discriminability (Item-Total Correlation [ITC])	Correlation between a correct response to an item and the overall score on the assessment. Positive numbers indicate that people who do well on the item do well on the overall assessment. Negative numbers indicate that people who do poorly on the item do well on the assessment. In this sense, ITC is both a measure of internal consistency and discriminability.
Effect Size (Cohen’s *d*)	Provides a size of the difference in the mean between the majority group and a specified minority group in standard deviation units (e.g., *d* = .5 indicates a half standard deviation difference in the means of two groups). In this report, *d* is expressed as a positive number and it is assumed that the majority group outperformed the minority group unless otherwise specified.
Omission Rates	Omission rates refer to the percentage of test-takers who did not provide a response to a given item. Item omission may be indicative of difficulty, IER, or speededness of the assessment.
Distractor Analysis	Distractor analysis refers to the percentage of test-takers who selected each response option. Effective distractors should attract at least some low-ability test-takers. If a distractor is selected more frequently than the correct response, this is a *potential* indicator of a poor distractor or a poor item.
Item Drift	Item drift refers to a change in an item’s parameters over time which may indicate a breach of test security.
Doublet Longstring	In this context, doublet longstring refers to test-takers who select the same response for the ME item and for the LE item for a single situation. This is indicative of IER or potentially speededness near the end of the assessment because it is impossible for the same action to be both the most effective and the least effective response to a single situation.
Flesch-Kincaid Reading Grade Level (RGL)	A readability formula used to assess approximate grade-level required to comprehend a selection of text. Lengthier sentences with longer words require more cognitive effort to comprehend and are assessed as higher grade level. The target reading level is 8th grade to ensure at least 80% of Americans can comprehend the selected text.
Principal Components Analysis (PCA)	A technique used to determine relations between items and factors. PCA assumes that there is both unique and shared variance among variables.
Paired Samples *t*-Test	A statistical procedure used in a repeated measures design where each test-taker is measured at all measurement occasions.
Independent Samples *t*-Test	A statistical procedure in which test-takers are assigned to one of two conditions.
Regression	A statistical technique which analyzes the statistical significance and relative importance of a predictor or collection of predictors in predicting a given outcome or outcomes.
R-Squared (R^2)	Indicates the proportion of variance in an outcome accounted for by a predictor or collection of predictors in a regression context.
Significance Testing	In this context, indicates the statistical significance of a given statistical test at a predetermined probability level. Indicates the probability that the results are due to chance characteristics alone rather than due to an actual effect. In this report, a traditional *p*-value of .05 was chosen for all statistical tests (i.e., $p \leq .05$) except where otherwise noted.
Stability Analysis	A specific case of the independent samples *t*-test. Stability analysis measures the change in test scores over time and is expressed using an effect size (Cohen’s *d*). In this context, stability analysis is used to examine potential test security breach.

APPENDIX D. Three Expert-Based Scoring Methods

Resp	Dichotomous										Trichotomous										Continuous									
	T1					T2					T1					T2					T1					T2				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
1	0	0	0	1	1	1	0	0	0	0	1	2	2	3	3	3	2	2	2	1	1	3	2	5	4	5	2	3	4	1
2	1	0	0	0	0	0	0	0	0	1	3	2	2	1	1	1	2	2	2	3	5	3	4	1	2	1	4	3	2	5
3	1	0	0	1	0	0	0	1	0	0	3	1	2	3	1	1	1	3	1	2	5	2	3	4	1	2	3	5	1	4
4	0	1	0	0	1	1	1	0	1	0	1	3	2	1	3	3	3	1	3	2	1	4	3	2	5	4	3	1	5	2
5	1	0	0	0	0	0	0	0	1	1	3	2	2	2	1	1	2	2	3	3	5	2	3	4	1	1	2	3	5	4
6	0	0	0	0	1	1	0	0	0	0	1	2	2	2	3	3	2	2	1	1	1	4	3	2	5	5	4	3	1	2
7	0	0	1	0	0	0	1	0	1	0	1	1	3	1	2	1	3	1	3	2	2	3	5	1	4	2	5	1	4	3
8	1	1	0	1	0	1	0	1	0	0	3	3	1	3	2	3	1	3	1	2	4	3	1	5	2	4	1	5	2	3
9	0	0	0	1	1	0	0	0	0	1	1	2	2	3	3	1	1	2	2	3	1	2	3	5	4	2	1	3	4	5
10	1	0	0	0	0	1	1	0	0	0	3	2	2	1	1	3	3	2	2	1	5	4	3	1	2	4	5	3	2	1
11	0	0	1	0	0	0	1	1	0	0	1	1	3	2	2	1	3	3	1	1	1	2	5	3	4	3	5	4	2	1
12	1	1	0	0	0	1	0	0	1	1	3	3	1	2	2	3	1	1	3	3	5	4	1	3	2	3	1	2	4	5
13	0	1	0	1	1	0	0	0	1	1	1	3	1	3	3	1	1	1	3	3	1	4	2	5	3	3	2	1	5	4
14	1	0	1	0	0	1	1	1	0	0	3	1	3	1	1	3	3	3	1	1	5	2	4	1	3	3	4	5	1	2
15	0	1	1	0	0	1	0	0	1	1	1	3	3	1	1	3	2	2	3	3	3	5	4	2	1	3	2	1	5	4
16	1	0	0	1	1	0	0	0	0	0	3	1	1	3	3	1	2	2	1	1	3	1	2	4	5	3	4	5	1	2
17	1	0	0	1	1	0	0	0	1	1	3	2	2	3	3	2	2	1	3	3	3	2	1	5	4	2	3	1	5	4
18	0	0	0	0	0	0	0	1	0	0	1	2	2	1	1	2	2	3	1	1	3	4	5	1	2	4	3	5	1	2
19	0	0	0	1	0	1	1	1	0	0	2	1	1	3	1	3	3	3	1	2	4	1	3	5	2	3	4	5	1	2
20	0	1	1	0	1	0	0	0	1	0	2	3	3	1	3	1	1	1	3	2	2	5	3	1	4	3	2	1	5	4
21	0	1	1	0	0	0	1	1	0	0	1	3	3	1	1	2	3	3	2	1	1	4	5	2	3	2	4	5	3	1
22	1	0	0	1	1	0	0	0	0	1	3	1	1	3	3	2	1	1	2	3	5	2	1	4	3	4	2	1	3	5
23	1	1	1	0	0	0	1	1	1	0	3	3	3	1	2	1	3	3	3	2	3	4	5	1	2	1	3	5	4	2
24	0	0	0	1	0	1	0	0	0	0	1	1	1	3	2	3	1	1	1	2	3	2	1	5	4	5	3	1	2	4
25	0	0	0	1	0	0	0	0	1	0	1	1	2	3	1	1	1	2	3	1	1	2	4	5	3	1	2	4	5	3
26	1	1	0	0	1	1	1	0	0	1	3	3	2	1	3	3	3	2	1	3	5	4	2	1	3	5	4	2	1	3
27	0	1	0	0	0	0	1	0	1	1	1	3	2	1	1	1	3	2	3	3	1	5	4	3	2	1	4	2	5	3
28	1	0	0	1	1	1	0	0	0	0	3	1	2	3	3	3	1	2	1	1	5	1	2	3	4	5	2	4	1	3
29	0	0	0	1	0	0	1	0	0	0	1	2	1	3	1	1	3	2	2	2	2	4	3	5	1	1	5	3	2	4
30	1	0	1	0	1	1	0	0	0	0	3	2	3	1	3	3	1	2	2	2	4	2	3	1	5	5	1	3	4	2
31	0	1	0	1	1	1	0	0	0	0	1	3	2	3	3	3	1	2	1	2	1	4	2	5	3	5	2	4	1	3
32	1	0	0	0	0	0	1	0	1	0	3	1	2	1	1	1	3	2	3	2	5	2	4	1	3	1	4	2	5	3
33	0	0	1	0	0	0	0	1	1	0	1	1	3	2	2	1	2	3	3	1	2	1	5	4	3	2	3	5	4	1
34	1	1	0	0	0	1	0	0	0	1	3	3	1	2	2	3	2	1	1	3	4	5	1	2	3	4	3	1	2	5
35	1	1	1	0	0	0	0	1	0	0	3	3	3	1	1	1	1	3	2	2	3	4	5	2	1	2	1	5	4	3
36	0	0	0	1	1	1	1	0	0	0	1	1	1	3	3	3	3	1	2	2	3	2	1	4	5	4	5	1	2	3
37	1	0	0	1	0	0	0	1	0	0	3	2	1	3	1	1	2	3	1	1	5	3	1	4	2	2	4	5	1	3
38	0	0	1	0	1	1	0	0	1	1	1	2	3	1	3	3	2	1	3	3	1	3	5	2	4	4	2	1	5	3
39	0	0	1	0	1	1	0	0	1	0	1	2	3	1	3	3	2	1	3	1	2	3	5	1	4	5	3	1	4	2
40	1	0	0	1	0	0	0	1	0	1	3	2	1	3	1	1	2	3	1	3	4	3	1	5	2	1	3	5	2	4
41	0	0	1	1	0	0	1	1	0	1	2	1	3	3	2	1	3	3	2	3	2	1	4	5	3	1	3	4	2	5
42	0	1	0	0	0	1	0	0	0	0	2	3	1	1	2	3	1	1	2	1	4	5	2	1	3	5	3	2	4	1
43	0	1	1	0	1	0	0	1	0	1	1	3	3	2	3	1	1	3	1	3	1	3	4	2	5	2	1	5	3	4

44	1	0	0	0	0	1	1	0	1	0	3	1	1	2	1	3	3	1	3	1	5	3	2	4	1	4	5	1	3	2
45	0	0	1	0	1	1	1	0	0	0	1	1	3	1	3	3	3	1	2	1	2	1	5	3	4	4	5	2	3	1
46	1	1	0	1	0	0	0	1	0	1	3	3	1	3	1	1	1	3	2	3	4	5	1	3	2	2	1	4	3	5
47	0	1	1	0	0	0	0	1	0	0	1	3	3	2	1	1	2	3	1	1	2	5	4	3	1	1	4	5	3	2
48	1	0	0	0	1	1	0	0	1	1	3	1	1	2	3	3	2	1	3	3	4	1	2	3	5	5	2	1	3	4
49	1	1	0	0	0	0	0	0	1	0	3	3	1	2	1	1	1	1	3	1	4	5	2	3	1	2	3	4	5	1
50	0	0	1	0	1	1	1	1	0	1	1	1	3	2	3	3	3	3	1	3	2	1	4	3	5	4	3	2	1	5

Note. Resp. = Response.

APPENDIX E. Stratified Samples Item-Level Descriptives

OTS-CIV

Item	Response	T1					T2				
		N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
1	1	8,014	2.93	.26	-3.76	13.60	7,274	2.82	.39	-1.85	1.90
	2	8,019	2.93	.32	-4.76	23.04	7,275	2.92	.29	-3.33	10.53
2	3	8,011	2.77	.46	-1.77	2.23	7,264	2.21	.81	-.39	-1.36
	4	8,013	2.94	.30	-5.67	32.19	7,274	2.91	.38	-4.38	18.25
3	5	8,014	2.82	.39	-1.82	1.78	7,275	2.79	.50	-2.32	4.51
	6	8,017	2.92	.29	-3.48	11.89	7,274	2.30	.74	-.55	-1.01
4	7	8,013	2.28	.76	-.51	-1.11	7,273	2.51	.80	-1.19	-.39
	8	8,009	2.93	.34	-4.98	24.19	7,280	2.58	.66	-1.30	.42
5	9	8,013	2.81	.47	-2.46	5.33	7,266	2.29	.60	-.23	-.61
	10	8,014	2.35	.72	-.64	-.86	7,274	2.78	.53	-2.34	4.42
6	11	8,015	2.79	.45	-1.99	3.20	7,275	2.38	.92	-.82	-1.32
	12	8,018	2.85	.36	-2.19	3.47	7,279	2.99	.15	-13.21	172.49
7	13	8,018	2.28	.96	-.58	-1.66	7,276	2.87	.50	-3.47	10.07
	14	8,016	1.98	1.00	.04	-2.00	7,278	2.93	.37	-4.98	22.81
8	15	8,014	2.47	.88	-1.05	-.89	7,271	2.86	.35	-2.04	2.17
	16	8,016	2.98	.18	-10.97	118.30	7,276	1.59	.49	-.36	-1.87
9	17	8,011	2.86	.35	-2.05	2.19	7,273	2.81	.41	-1.82	2.07
	18	8,013	1.63	.48	-.55	-1.70	7,274	2.48	.71	-1.01	-.35
10	19	8,012	2.27	.83	-.53	-1.36	7,274	2.82	.40	-1.91	2.41
	20	8,016	2.93	.31	-4.58	21.65	7,277	2.86	.47	-3.25	9.29
11	21	8,009	2.67	.74	-1.81	1.26	7,272	2.51	.51	-.13	-1.78
	22	8,011	2.94	.34	-5.54	28.68	7,272	2.70	.58	-1.78	2.05
12	23	8,007	2.84	.39	-2.14	3.50	7,270	2.95	.29	-5.78	33.46
	24	8,010	2.85	.50	-3.17	8.55	7,277	2.81	.47	-2.54	5.71
13	25	8,005	2.28	.93	-.59	-1.58	7,256	2.23	.95	-.48	-1.71
	26	8,008	2.90	.35	-3.86	15.08	7,262	2.89	.38	-3.54	12.45
14	27	8,010	2.31	.84	-.63	-1.28	7,263	2.85	.39	-2.44	5.31
	28	8,009	2.99	.13	-13.78	198.89	7,261	2.72	.67	-2.06	2.39
15	29	8,006	2.41	.86	-.89	-1.06	7,244	2.21	.62	-.18	-.59
	30	8,003	2.96	.24	-6.20	40.87	7,245	2.31	.66	-.42	-.74
16	31	7,987	2.79	.44	-1.88	2.62	7,236	2.75	.51	-1.89	2.70
	32	7,991	2.64	.75	-1.64	.82	7,242	2.78	.46	-1.89	2.75
17	33	7,989	2.83	.45	-2.79	7.09	7,222	2.84	.41	-2.51	5.83
	34	7,991	2.90	.33	-3.38	11.52	7,222	2.96	.26	-7.03	48.94
18	35	7,964	2.85	.52	-3.25	8.55	7,185	2.80	.50	-2.48	5.28
	36	7,961	2.78	.63	-2.47	4.10	7,185	2.89	.35	-3.34	11.27
19	37	7,909	2.38	.82	-.79	-1.05	7,141	2.10	.84	-.19	-1.57
	38	7,912	2.32	.86	-.67	-1.32	7,139	2.94	.31	-5.55	30.34
20	39	7,874	2.70	.63	-1.93	2.27	7,088	2.37	.83	-.78	-1.09
	40	7,868	2.80	.54	-2.65	5.66	7,085	2.32	.84	-.66	-1.25
21	41	7,814	2.54	.54	-.55	-.88	7,045	2.91	.36	-4.27	17.87
	42	7,794	2.31	.75	-.58	-1.04	7,044	2.60	.76	-1.48	.38
22	43	7,743	2.91	.35	-4.31	18.49	6,963	2.80	.60	-2.65	5.00
	44	7,734	2.59	.77	-1.45	.27	6,962	2.86	.51	-3.40	9.54
23	45	7,654	2.79	.62	-2.54	4.47	6,848	2.78	.57	-2.44	4.44
	46	7,650	2.86	.51	-3.36	9.27	6,861	2.81	.55	-2.77	6.04
24	47	7,570	2.02	.91	-.05	-1.78	6,724	2.62	.71	-1.56	.80
	48	7,554	2.10	.93	-.19	-1.82	6,717	2.86	.45	-3.26	9.61
25	49	7,474	2.79	.57	-2.54	4.91	6,697	2.32	.95	-.67	-1.55
	50	7,492	2.82	.54	-2.85	6.45	6,684	2.93	.36	-5.25	25.55

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

OTS-AD

Item	Response	T1					T2				
		N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
1	1	5,426	2.95	.23	-4.51	20.48	4,834	2.81	.39	-1.64	.77
	2	5,426	2.96	.26	-6.43	42.37	4,836	2.94	.24	-4.26	18.39
2	3	5,421	2.74	.50	-1.73	2.13	4,831	2.30	.80	-.60	-1.20
	4	5,425	2.92	.36	-4.81	21.83	4,831	2.94	.31	-5.58	30.61
3	5	5,420	2.79	.41	-1.53	.63	4,835	2.88	.36	-3.10	9.54
	6	5,422	2.94	.25	-4.06	16.49	4,833	2.47	.66	-.88	-.37
4	7	5,422	2.38	.73	-.71	-.81	4,834	2.58	.78	-1.40	.12
	8	5,426	2.97	.24	-7.29	53.72	4,836	2.73	.51	-1.73	2.12
5	9	5,424	2.89	.34	-3.19	10.13	4,832	2.50	.59	-.71	-.47
	10	5,428	2.46	.66	-.85	-.41	4,836	2.88	.38	-3.33	11.00
6	11	5,426	2.84	.38	-2.15	3.46	4,835	2.42	.91	-.92	-1.16
	12	5,427	2.94	.24	-3.98	15.18	4,835	2.99	.14	-14.08	196.38
7	13	5,427	2.39	.92	-.86	-1.26	4,832	2.91	.41	-4.45	17.83
	14	5,428	2.06	1.00	-.13	-1.98	4,836	2.97	.25	-7.89	60.34
8	15	5,425	2.50	.86	-1.16	-.64	4,834	2.85	.36	-1.97	1.90
	16	5,428	2.99	.14	-14.63	212.05	4,836	1.52	.50	-.07	-2.00
9	17	5,423	2.82	.38	-1.66	.77	4,832	2.83	.38	-1.96	2.48
	18	5,425	1.55	.50	-.20	-1.96	4,834	2.57	.66	-1.27	.34
10	19	5,420	2.35	.89	-.74	-1.32	4,833	2.87	.35	-2.47	5.10
	20	5,425	2.92	.30	-4.07	17.32	4,835	2.90	.39	-4.00	15.41
11	21	5,420	2.68	.73	-1.87	1.49	4,833	2.50	.51	-.11	-1.75
	22	5,424	2.96	.30	-6.47	39.83	4,835	2.71	.52	-1.63	1.74
12	23	5,426	2.87	.35	-2.47	5.11	4,833	2.97	.23	-7.40	56.52
	24	5,423	2.90	.40	-4.20	16.47	4,835	2.84	.44	-2.88	7.74
13	25	5,424	2.31	.93	-.65	-1.52	4,826	2.26	.94	-.54	-1.66
	26	5,418	2.89	.36	-3.36	11.32	4,832	2.89	.36	-3.35	11.27
14	27	5,423	2.52	.78	-1.22	-.24	4,827	2.86	.37	-2.52	5.64
	28	5,425	2.99	.11	-16.85	294.34	4,830	2.82	.55	-2.82	6.19
15	29	5,417	2.51	.84	-1.19	-.52	4,821	2.28	.62	-.28	-.65
	30	5,415	2.97	.19	-7.55	62.31	4,824	2.28	.65	-.36	-.75
16	31	5,417	2.80	.43	-1.91	2.66	4,824	2.77	.51	-2.14	3.71
	32	5,416	2.73	.66	-2.13	2.67	4,824	2.82	.41	-2.02	3.15
17	33	5,409	2.85	.42	-2.87	7.82	4,814	2.77	.48	-1.93	2.93
	34	5,412	2.93	.28	-4.34	19.86	4,820	2.96	.27	-6.78	45.42
18	35	5,404	2.91	.41	-4.41	17.46	4,807	2.81	.48	-2.51	5.53
	36	5,405	2.87	.50	-3.48	10.08	4,805	2.92	.32	-4.06	17.08
19	37	5,389	2.45	.85	-.99	-.89	4,779	2.21	.84	-.42	-1.46
	38	5,385	2.33	.78	-.64	-1.08	4,789	2.95	.30	-5.78	33.02
20	39	5,356	2.79	.52	-2.51	5.16	4,752	2.38	.88	-.82	-1.21
	40	5,362	2.89	.42	-3.99	14.50	4,750	2.31	.75	-.58	-1.03
21	41	5,338	2.64	.51	-.93	-.37	4,732	2.94	.31	-5.43	29.35
	42	5,334	2.61	.64	-1.37	.67	4,733	2.75	.60	-2.24	3.46
22	43	5,309	2.95	.26	-6.14	39.03	4,713	2.82	.58	-2.82	5.93
	44	5,309	2.74	.62	-2.19	3.19	4,713	2.87	.49	-3.52	10.37
23	45	5,285	2.80	.60	-2.69	5.21	4,670	2.80	.55	-2.64	5.48
	46	5,282	2.87	.49	-3.54	10.51	4,674	2.82	.54	-2.91	6.79
24	47	5,241	2.20	.93	-.41	-1.71	4,629	2.64	.69	-1.63	1.04
	48	5,232	2.04	.92	-.08	-1.82	4,629	2.87	.44	-3.45	10.84
25	49	5,211	2.81	.53	-2.68	5.78	4,610	2.28	.96	-.59	-1.65
	50	5,213	2.85	.50	-3.17	8.46	4,603	2.94	.35	-5.30	26.12

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

AECP

Item	Response	T1					T2				
		N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
1	1	170	2.95	.21	-4.24	16.07	167	2.80	.40	-1.51	.27
	2	170	2.96	.21	-6.64	47.87	167	2.95	.25	-4.99	26.78
2	3	170	2.71	.49	-1.36	.79	168	2.23	.76	-.40	-1.17
	4	170	2.95	.31	-5.84	32.87	168	2.92	.38	-4.74	20.85
3	5	170	2.79	.41	-1.40	-.04	167	2.88	.34	-2.75	7.02
	6	170	2.95	.21	-4.24	16.07	168	2.46	.69	-.88	-.48
4	7	170	2.31	.75	-.57	-1.03	168	2.40	.90	-.88	-1.20
	8	170	2.93	.32	-4.79	23.30	168	2.67	.58	-1.59	1.44
5	9	170	2.89	.31	-2.54	4.47	168	2.45	.60	-.55	-.64
	10	170	2.42	.69	-.76	-.64	168	2.84	.41	-2.55	6.06
6	11	170	2.88	.34	-2.79	7.23	167	2.32	.95	-.66	-1.57
	12	170	2.94	.24	-3.72	11.89	168	2.98	.22	-8.92	78.04
7	13	170	2.41	.91	-.90	-1.20	168	2.89	.45	-3.93	13.52
	14	170	2.08	1.00	-.16	-1.98	168	2.92	.40	-4.55	18.78
8	15	170	2.47	.88	-1.06	-.89	168	2.85	.36	-1.89	1.59
	16	170	2.96	.26	-7.26	51.04	168	1.53	.50	-.12	-2.00
9	17	170	2.85	.36	-1.91	1.66	168	2.80	.41	-1.76	1.84
	18	170	1.49	.50	.02	-2.01	168	2.53	.68	-1.12	-.06
10	19	170	2.22	.93	-.45	-1.69	168	2.85	.36	-1.96	1.84
	20	170	2.89	.35	-3.22	10.41	168	2.85	.51	-3.13	8.13
11	21	170	2.58	.82	-1.40	-.04	168	2.51	.50	-.02	-2.01
	22	170	2.93	.37	-4.99	23.06	168	2.73	.57	-1.93	2.63
12	23	170	2.82	.39	-1.63	.66	168	2.92	.35	-4.42	19.27
	24	170	2.85	.51	-3.15	8.28	168	2.81	.49	-2.55	5.65
13	25	170	2.23	.95	-.47	-1.75	168	2.21	.97	-.43	-1.80
	26	170	2.87	.43	-3.38	10.72	168	2.88	.38	-3.15	9.83
14	27	170	2.48	.79	-1.05	-.60	168	2.83	.40	-2.30	4.62
	28	170	3.00	.00	NA	NA	168	2.79	.59	-2.51	4.58
15	29	170	2.43	.90	-.94	-1.11	168	2.30	.63	-.34	-.71
	30	170	2.96	.19	-4.99	23.06	168	2.23	.60	-.13	-.52
16	31	170	2.82	.38	-1.68	.84	168	2.79	.45	-2.00	3.22
	32	170	2.76	.62	-2.35	3.74	168	2.79	.42	-1.66	1.42
17	33	170	2.81	.46	-2.43	5.26	167	2.75	.49	-1.74	2.15
	34	170	2.97	.20	-7.49	60.47	167	2.96	.28	-6.61	42.93
18	35	170	2.94	.34	-5.52	28.65	167	2.71	.61	-1.94	2.39
	36	170	2.78	.63	-2.44	3.99	167	2.95	.25	-4.99	26.78
19	37	169	2.41	.87	-.90	-1.08	166	2.23	.84	-.45	-1.46
	38	169	2.41	.78	-.86	-.84	165	2.93	.36	-4.89	22.73
20	39	169	2.80	.47	-2.35	4.86	164	2.35	.89	-.75	-1.33
	40	169	2.91	.41	-4.22	16.27	163	2.32	.76	-.60	-1.05
21	41	169	2.63	.54	-1.07	.09	162	2.90	.41	-4.09	15.58
	42	169	2.63	.62	-1.43	.86	159	2.58	.75	-1.41	.26
22	43	168	2.97	.20	-7.44	59.70	158	2.82	.57	-2.87	6.26
	44	168	2.60	.77	-1.49	.35	158	2.92	.38	-4.79	21.07
23	45	166	2.76	.65	-2.31	3.36	156	2.74	.61	-2.17	3.17
	46	165	2.82	.58	-2.82	5.99	155	2.81	.56	-2.69	5.60
24	47	164	2.16	.94	-.33	-1.78	154	2.68	.65	-1.80	1.72
	48	165	1.98	.94	.04	-1.87	154	2.92	.36	-4.41	19.02
25	49	163	2.76	.58	-2.26	3.75	154	2.22	.98	-.45	-1.81
	50	164	2.85	.50	-3.24	8.88	154	2.95	.32	-5.90	33.05

Note. *N* = Sample Size; *M* = Mean; *SD* = Standard Deviation. NA indicates that the parameters could not be estimated due to lack of variability.

USAFA

Item	Response	T1					T2				
		N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
1	1	4,298	2.93	.26	-3.74	13.58	3,733	2.78	.42	-1.42	.21
	2	4,300	2.97	.22	-7.86	63.12	3,735	2.95	.23	-5.42	31.72
2	3	4,295	2.90	.34	-3.72	14.19	3,734	2.25	.83	-.50	-1.38
	4	4,297	2.97	.23	-7.55	58.52	3,736	2.91	.39	-4.20	16.71
3	5	4,299	2.76	.44	-1.36	.24	3,733	2.88	.40	-3.45	11.52
	6	4,301	2.95	.24	-5.44	31.87	3,735	2.49	.67	-.93	-.30
4	7	4,296	2.33	.79	-.64	-1.10	3,732	2.18	.97	-.36	-1.83
	8	4,298	2.92	.35	-4.61	20.77	3,735	2.66	.56	-1.40	.98
5	9	4,298	2.90	.36	-3.72	13.88	3,736	2.47	.57	-.49	-.75
	10	4,300	2.51	.66	-.99	-.19	3,735	2.90	.36	-3.74	14.03
6	11	4,298	2.89	.37	-3.67	13.39	3,733	2.53	.85	-1.25	-.43
	12	4,301	2.86	.36	-2.30	4.14	3,734	2.98	.18	-10.66	111.63
7	13	4,297	2.42	.91	-.92	-1.15	3,731	2.90	.44	-4.10	14.78
	14	4,299	2.02	1.00	-.03	-2.00	3,733	2.94	.33	-5.74	30.90
8	15	4,299	2.59	.81	-1.45	.09	3,733	2.86	.34	-2.11	2.46
	16	4,297	2.99	.16	-12.73	160.20	3,735	1.59	.49	-.37	-1.86
9	17	4,296	2.82	.38	-1.71	.92	3,735	2.85	.37	-2.20	3.57
	18	4,300	1.61	.49	-.47	-1.78	3,736	2.53	.66	-1.09	-.02
10	19	4,300	2.29	.91	-.60	-1.51	3,735	2.89	.33	-3.03	8.87
	20	4,300	2.93	.28	-4.53	21.73	3,735	2.81	.52	-2.71	6.08
11	21	4,294	2.64	.77	-1.65	.73	3,734	2.50	.51	-.05	-1.84
	22	4,298	2.94	.33	-5.65	29.96	3,735	2.68	.55	-1.51	1.32
12	23	4,302	2.91	.31	-3.67	13.86	3,733	2.95	.30	-6.18	36.76
	24	4,299	2.82	.52	-2.80	6.49	3,730	2.77	.48	-1.97	3.09
13	25	4,298	2.27	.93	-.57	-1.60	3,735	2.25	.94	-.51	-1.68
	26	4,298	2.91	.33	-3.77	14.60	3,733	2.90	.34	-3.59	13.17
14	27	4,297	2.58	.70	-1.35	.32	3,731	2.91	.32	-3.55	12.85
	28	4,302	2.99	.16	-11.42	134.46	3,734	2.75	.63	-2.25	3.29
15	29	4,301	2.62	.75	-1.58	.63	3,728	2.12	.61	-.08	-.42
	30	4,302	2.95	.25	-5.71	34.88	3,733	2.19	.72	-.29	-1.03
16	31	4,297	2.86	.38	-2.70	6.87	3,733	2.81	.48	-2.52	5.52
	32	4,298	2.66	.72	-1.77	1.28	3,731	2.67	.49	-.97	-.48
17	33	4,299	2.83	.45	-2.68	6.58	3,727	2.88	.36	-3.12	9.62
	34	4,299	2.94	.27	-4.99	26.48	3,728	2.98	.20	-9.34	87.47
18	35	4,291	2.92	.40	-4.60	19.14	3,723	2.82	.46	-2.61	6.17
	36	4,294	2.83	.56	-2.96	6.79	3,726	2.94	.27	-4.91	25.68
19	37	4,285	2.61	.70	-1.52	.72	3,715	2.32	.83	-.66	-1.23
	38	4,291	2.55	.71	-1.26	.10	3,715	2.96	.25	-7.15	50.72
20	39	4,280	2.79	.50	-2.36	4.69	3,706	2.63	.70	-1.57	.85
	40	4,284	2.88	.46	-3.63	11.60	3,711	2.53	.69	-1.14	-.03
21	41	4,271	2.76	.45	-1.55	1.21	3,699	2.95	.28	-5.64	32.78
	42	4,270	2.61	.62	-1.35	.68	3,700	2.63	.73	-1.60	.76
22	43	4,266	2.93	.31	-4.96	25.07	3,687	2.80	.59	-2.71	5.35
	44	4,263	2.63	.73	-1.61	.82	3,687	2.89	.46	-3.89	13.12
23	45	4,250	2.80	.60	-2.70	5.27	3,668	2.81	.53	-2.67	5.77
	46	4,248	2.88	.48	-3.63	11.21	3,666	2.85	.51	-3.15	8.22
24	47	4,230	1.92	.95	.16	-1.87	3,637	2.78	.57	-2.42	4.37
	48	4,229	1.87	.91	.25	-1.75	3,640	2.91	.37	-4.22	17.46
25	49	4,193	2.80	.53	-2.64	5.65	3,622	2.43	.90	-.95	-1.10
	50	4,197	2.85	.51	-3.23	8.69	3,620	2.97	.23	-8.33	67.38

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

ROTC

Item	Response	T1					T2				
		N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
1	1	12,750	2.93	.27	-3.85	14.98	12,392	2.79	.41	-1.47	.37
	2	12,758	2.95	.28	-5.83	34.57	12,399	2.94	.25	-4.51	21.37
2	3	12,744	2.87	.38	-3.16	9.83	12,375	2.06	.85	-.11	-1.60
	4	12,751	2.95	.28	-5.86	34.60	12,383	2.89	.42	-3.96	14.47
3	5	12,738	2.76	.43	-1.34	.07	12,388	2.66	.62	-1.67	1.49
	6	12,748	2.94	.25	-4.61	22.57	12,388	2.19	.79	-.36	-1.31
4	7	12,737	2.16	.81	-.30	-1.42	12,385	2.32	.92	-.68	-1.46
	8	12,747	2.90	.40	-4.10	15.67	12,392	2.59	.61	-1.20	.37
5	9	12,741	2.68	.61	-1.74	1.76	12,390	2.34	.61	-.32	-.66
	10	12,741	2.23	.79	-.42	-1.27	12,398	2.83	.45	-2.73	6.83
6	11	12,750	2.79	.44	-1.90	2.69	12,384	2.22	.97	-.46	-1.79
	12	12,754	2.90	.31	-2.91	7.56	12,393	2.98	.19	-10.47	107.65
7	13	12,735	2.31	.95	-.64	-1.59	12,390	2.83	.56	-2.94	6.63
	14	12,739	1.92	1.00	.17	-1.97	12,394	2.91	.42	-4.37	17.06
8	15	12,741	2.34	.94	-.73	-1.46	12,379	2.79	.41	-1.41	-.02
	16	12,749	2.98	.18	-10.67	111.95	12,390	1.54	.50	-.18	-1.97
9	17	12,734	2.73	.45	-1.01	-.97	12,391	2.82	.40	-2.04	3.13
	18	12,740	1.56	.50	-.26	-1.93	12,398	2.46	.72	-.94	-.48
10	19	12,743	2.22	.89	-.45	-1.60	12,380	2.82	.40	-2.03	3.13
	20	12,751	2.91	.33	-4.11	17.27	12,395	2.84	.49	-3.09	8.21
11	21	12,737	2.60	.80	-1.50	.26	12,388	2.45	.51	.08	-1.68
	22	12,740	2.92	.39	-4.67	19.82	12,394	2.68	.55	-1.51	1.33
12	23	12,734	2.83	.40	-2.27	4.44	12,384	2.95	.30	-6.04	35.58
	24	12,739	2.83	.52	-2.93	7.05	12,387	2.76	.51	-2.08	3.46
13	25	12,723	2.21	.95	-.42	-1.75	12,364	2.14	.96	-.28	-1.86
	26	12,719	2.89	.37	-3.57	12.74	12,367	2.88	.38	-3.40	11.49
14	27	12,721	2.44	.82	-.97	-.80	12,368	2.86	.37	-2.74	7.13
	28	12,733	2.98	.17	-10.57	115.41	12,371	2.73	.66	-2.11	2.65
15	29	12,712	2.45	.86	-1.00	-.91	12,328	2.18	.69	-.26	-.91
	30	12,709	2.95	.25	-5.40	31.25	12,335	2.19	.64	-.20	-.67
16	31	12,690	2.82	.42	-2.16	3.89	12,334	2.76	.52	-2.12	3.54
	32	12,698	2.62	.75	-1.57	.61	12,333	2.71	.48	-1.34	.69
17	33	12,675	2.80	.49	-2.40	4.94	12,295	2.85	.40	-2.85	7.78
	34	12,679	2.91	.32	-3.86	15.40	12,310	2.96	.27	-6.64	43.36
18	35	12,648	2.79	.61	-2.61	4.80	12,276	2.76	.53	-2.17	3.71
	36	12,647	2.71	.70	-2.04	2.15	12,275	2.90	.35	-3.68	13.75
19	37	12,606	2.37	.84	-.79	-1.13	12,220	2.18	.83	-.34	-1.48
	38	12,591	2.35	.83	-.73	-1.16	12,218	2.94	.31	-5.58	30.31
20	39	12,548	2.72	.58	-1.94	2.54	12,114	2.40	.84	-.87	-1.02
	40	12,547	2.83	.52	-2.98	7.38	12,115	2.36	.80	-.75	-1.03
21	41	12,470	2.65	.52	-1.05	-.03	12,044	2.90	.38	-4.11	16.39
	42	12,456	2.41	.73	-.81	-.70	12,040	2.56	.78	-1.35	.01
22	43	12,390	2.90	.37	-4.03	15.91	11,948	2.75	.66	-2.26	3.11
	44	12,379	2.54	.80	-1.28	-.20	11,941	2.84	.54	-3.10	7.63
23	45	12,283	2.74	.67	-2.20	2.85	11,808	2.71	.64	-1.99	2.37
	46	12,268	2.84	.54	-3.11	7.65	11,825	2.75	.64	-2.25	3.27
24	47	12,185	2.02	.91	-.04	-1.79	11,678	2.62	.71	-1.55	.75
	48	12,157	1.96	.93	.08	-1.85	11,679	2.85	.46	-3.13	8.80
25	49	12,045	2.73	.62	-2.08	2.79	11,634	2.29	.96	-.62	-1.62
	50	12,053	2.74	.64	-2.20	3.04	11,618	2.94	.34	-5.46	27.85

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

ANG

Item	Response	T1					T2				
		N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
1	1	3,913	2.94	.25	-4.10	16.70	3,287	2.82	.39	-1.67	.91
	2	3,915	2.96	.23	-6.89	49.85	3,288	2.94	.24	-4.25	18.04
2	3	3,912	2.79	.45	-2.06	3.52	3,284	2.32	.79	-.63	-1.13
	4	3,912	2.93	.34	-5.05	24.70	3,287	2.94	.31	-5.61	30.79
3	5	3,912	2.78	.42	-1.41	.12	3,286	2.85	.40	-2.71	6.95
	6	3,913	2.94	.26	-4.16	17.74	3,285	2.49	.69	-.98	-.32
4	7	3,910	2.40	.72	-.78	-.71	3,284	2.50	.83	-1.14	-.57
	8	3,910	2.97	.22	-7.67	61.11	3,285	2.67	.57	-1.58	1.46
5	9	3,911	2.87	.38	-2.91	8.21	3,284	2.40	.60	-.47	-.65
	10	3,913	2.50	.68	-1.02	-.21	3,286	2.85	.42	-2.85	7.69
6	11	3,913	2.82	.41	-2.11	3.60	3,284	2.46	.89	-1.03	-.94
	12	3,912	2.93	.27	-3.54	11.79	3,287	2.98	.18	-11.10	121.36
7	13	3,909	2.46	.89	-1.03	-.94	3,286	2.91	.42	-4.34	16.80
	14	3,912	2.10	.99	-.21	-1.96	3,286	2.96	.29	-6.48	40.01
8	15	3,912	2.54	.84	-1.27	-.39	3,284	2.87	.33	-2.26	3.11
	16	3,911	2.99	.15	-13.53	181.15	3,287	1.59	.49	-.37	-1.87
9	17	3,912	2.86	.35	-2.06	2.24	3,286	2.79	.42	-1.73	1.79
	18	3,913	1.61	.49	-.47	-1.78	3,285	2.56	.67	-1.23	.22
10	19	3,910	2.32	.88	-.67	-1.37	3,284	2.86	.36	-2.39	4.73
	20	3,914	2.92	.30	-4.20	18.49	3,283	2.88	.43	-3.64	12.20
11	21	3,908	2.65	.76	-1.70	.90	3,283	2.53	.50	-.20	-1.79
	22	3,910	2.95	.30	-6.36	38.45	3,282	2.72	.52	-1.70	1.99
12	23	3,913	2.87	.35	-2.46	5.00	3,285	2.97	.24	-7.38	54.97
	24	3,915	2.87	.46	-3.55	11.19	3,284	2.82	.47	-2.57	5.89
13	25	3,910	2.25	.94	-.52	-1.66	3,281	2.22	.95	-.44	-1.75
	26	3,911	2.88	.38	-3.27	10.62	3,281	2.87	.38	-3.11	9.49
14	27	3,910	2.48	.78	-1.08	-.52	3,278	2.85	.38	-2.41	5.04
	28	3,912	2.99	.15	-12.33	154.37	3,280	2.77	.62	-2.39	3.89
15	29	3,912	2.50	.83	-1.15	-.56	3,278	2.28	.61	-.25	-.62
	30	3,912	2.97	.20	-6.85	51.14	3,279	2.34	.65	-.46	-.70
16	31	3,903	2.81	.42	-2.04	3.29	3,274	2.79	.47	-2.21	4.18
	32	3,908	2.69	.70	-1.89	1.70	3,274	2.81	.42	-2.03	3.26
17	33	3,899	2.82	.46	-2.56	5.89	3,269	2.80	.44	-2.04	3.38
	34	3,904	2.93	.29	-4.30	19.49	3,273	2.96	.28	-6.68	43.51
18	35	3,896	2.90	.44	-4.09	14.72	3,264	2.81	.47	-2.48	5.46
	36	3,896	2.85	.53	-3.18	8.11	3,265	2.91	.33	-4.12	17.46
19	37	3,881	2.43	.84	-.95	-.91	3,255	2.20	.84	-.39	-1.48
	38	3,880	2.33	.81	-.67	-1.15	3,261	2.94	.31	-5.70	31.88
20	39	3,876	2.78	.53	-2.37	4.54	3,248	2.40	.86	-.88	-1.06
	40	3,880	2.86	.48	-3.38	9.85	3,252	2.33	.77	-.64	-1.04
21	41	3,863	2.68	.51	-1.17	.23	3,243	2.95	.29	-5.69	32.47
	42	3,863	2.43	.70	-.84	-.57	3,238	2.71	.66	-1.98	2.21
22	43	3,848	2.95	.29	-5.73	32.99	3,211	2.84	.54	-3.08	7.48
	44	3,845	2.71	.65	-1.99	2.30	3,217	2.89	.45	-3.93	13.43
23	45	3,821	2.82	.57	-2.86	6.20	3,179	2.81	.54	-2.68	5.76
	46	3,824	2.87	.49	-3.52	10.36	3,182	2.83	.53	-2.99	7.23
24	47	3,798	2.07	.94	-.14	-1.86	3,155	2.66	.68	-1.73	1.38
	48	3,798	2.06	.92	-.11	-1.82	3,152	2.86	.47	-3.29	9.63
25	49	3,773	2.82	.51	-2.84	6.79	3,142	2.35	.94	-.74	-1.45
	50	3,773	2.85	.50	-3.16	8.37	3,138	2.95	.32	-5.94	33.27

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

Item	Response	T1					T2				
		N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
1	1	1,481	2.94	.26	-4.03	16.36	1,215	2.81	.39	-1.64	.82
	2	1,480	2.95	.28	-5.76	33.96	1,215	2.94	.25	-4.03	16.25
2	3	1,481	2.77	.47	-1.92	2.90	1,214	2.31	.81	-.63	-1.18
	4	1,480	2.94	.32	-5.39	28.28	1,214	2.93	.33	-5.10	25.33
3	5	1,479	2.77	.42	-1.33	-.14	1,214	2.86	.39	-2.91	8.23
	6	1,479	2.94	.24	-4.38	19.75	1,215	2.45	.70	-.88	-.52
4	7	1,478	2.37	.73	-.71	-.83	1,214	2.50	.84	-1.14	-.61
	8	1,479	2.95	.29	-5.71	32.77	1,215	2.66	.57	-1.44	1.08
5	9	1,480	2.87	.39	-3.04	9.02	1,214	2.43	.60	-.52	-.64
	10	1,481	2.48	.68	-.94	-.36	1,215	2.86	.42	-3.05	8.85
6	11	1,481	2.81	.42	-1.92	2.65	1,215	2.42	.91	-.93	-1.13
	12	1,481	2.93	.26	-3.28	8.78	1,215	2.99	.11	-17.32	298.26
7	13	1,480	2.44	.90	-.99	-1.02	1,213	2.89	.45	-3.96	13.69
	14	1,478	2.04	1.00	-.07	-2.00	1,215	2.94	.34	-5.54	28.73
8	15	1,480	2.51	.86	-1.18	-.60	1,214	2.87	.34	-2.16	2.68
	16	1,480	2.99	.16	-12.03	142.81	1,214	1.58	.49	-.31	-1.90
9	17	1,478	2.86	.35	-2.08	2.33	1,212	2.78	.44	-1.62	1.40
	18	1,479	1.60	.49	-.40	-1.84	1,214	2.54	.69	-1.19	.04
10	19	1,480	2.29	.88	-.61	-1.43	1,214	2.85	.37	-2.25	3.87
	20	1,481	2.90	.34	-3.68	13.80	1,214	2.87	.44	-3.55	11.48
11	21	1,481	2.68	.73	-1.85	1.43	1,213	2.53	.51	-.18	-1.77
	22	1,481	2.94	.34	-5.60	29.43	1,214	2.71	.54	-1.65	1.81
12	23	1,480	2.84	.39	-2.29	4.48	1,213	2.97	.22	-7.47	58.02
	24	1,481	2.86	.46	-3.37	10.16	1,213	2.81	.48	-2.57	5.84
13	25	1,480	2.30	.93	-.64	-1.53	1,215	2.23	.95	-.48	-1.73
	26	1,480	2.88	.38	-3.36	11.13	1,215	2.89	.38	-3.52	12.30
14	27	1,478	2.45	.80	-1.00	-.69	1,214	2.83	.40	-2.26	4.34
	28	1,479	2.99	.15	-12.62	162.40	1,215	2.76	.64	-2.33	3.52
15	29	1,477	2.48	.85	-1.08	-.74	1,211	2.28	.61	-.24	-.62
	30	1,477	2.96	.25	-6.26	41.02	1,210	2.35	.64	-.47	-.69
16	31	1,474	2.81	.42	-2.11	3.64	1,209	2.76	.50	-2.04	3.35
	32	1,476	2.67	.72	-1.78	1.27	1,209	2.81	.42	-2.13	3.76
17	33	1,473	2.80	.49	-2.40	4.94	1,209	2.80	.45	-2.10	3.68
	34	1,473	2.92	.31	-4.09	17.40	1,209	2.97	.24	-7.74	59.83
18	35	1,468	2.91	.42	-4.35	16.93	1,209	2.80	.46	-2.35	4.85
	36	1,469	2.86	.51	-3.36	9.32	1,206	2.93	.31	-4.47	20.81
19	37	1,463	2.38	.87	-.81	-1.19	1,198	2.16	.84	-.31	-1.51
	38	1,464	2.34	.80	-.68	-1.12	1,200	2.94	.33	-5.27	26.84
20	39	1,459	2.77	.55	-2.32	4.16	1,192	2.34	.89	-.71	-1.35
	40	1,462	2.88	.45	-3.64	11.75	1,189	2.33	.79	-.64	-1.10
21	41	1,456	2.64	.53	-1.06	.04	1,185	2.93	.33	-5.12	25.73
	42	1,454	2.49	.69	-1.02	-.27	1,183	2.68	.69	-1.83	1.61
22	43	1,449	2.95	.28	-5.81	34.65	1,174	2.80	.61	-2.62	4.88
	44	1,450	2.70	.65	-1.96	2.19	1,171	2.87	.50	-3.44	9.87
23	45	1,430	2.81	.58	-2.80	5.84	1,156	2.79	.55	-2.53	5.01
	46	1,430	2.89	.46	-3.89	13.14	1,158	2.83	.52	-2.99	7.41
24	47	1,421	2.12	.93	-.23	-1.80	1,154	2.66	.67	-1.70	1.32
	48	1,418	2.12	.93	-.25	-1.79	1,152	2.84	.49	-3.10	8.25
25	49	1,420	2.83	.50	-2.90	7.20	1,153	2.34	.94	-.72	-1.49
	50	1,420	2.85	.50	-3.20	8.66	1,151	2.95	.30	-6.29	37.58

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

APPENDIX F. Stratified Samples Item-Level Difficulty (p-values)

Item	Response	OTS-CIV		OTS-AD		AECP		USAFA		ROTC		ANG		AFRES	
		T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
1	1	.93	.83	.95	.82	.95	.80	.93	.78	.93	.79	.94	.82	.94	.81
	2	.95	.92	.97	.94	.97	.95	.98	.96	.96	.95	.97	.95	.96	.94
2	3	.78	.45	.77	.52	.73	.42	.92	.51	.89	.39	.81	.52	.80	.53
	4	.96	.94	.95	.96	.97	.96	.98	.94	.96	.93	.96	.96	.96	.96
3	5	.83	.83	.80	.89	.79	.89	.76	.91	.77	.75	.78	.87	.77	.88
	6	.92	.47	.94	.57	.95	.57	.96	.58	.95	.43	.94	.60	.95	.57
4	7	.47	.71	.52	.76	.48	.69	.53	.57	.42	.63	.54	.71	.53	.72
	8	.95	.68	.98	.76	.95	.73	.95	.70	.94	.65	.98	.73	.96	.71
5	9	.84	.37	.90	.55	.89	.50	.92	.51	.76	.41	.88	.47	.89	.49
	10	.50	.83	.56	.90	.54	.86	.60	.92	.45	.86	.61	.87	.59	.88
6	11	.81	.69	.85	.71	.89	.66	.91	.77	.81	.61	.83	.73	.82	.71
	12	.86	.99	.94	1.00	.94	.99	.86	.99	.90	.99	.93	.99	.93	1.00
7	13	.64	.93	.70	.96	.71	.95	.71	.95	.65	.91	.73	.95	.72	.95
	14	.49	.96	.53	.98	.54	.96	.51	.97	.46	.95	.55	.98	.52	.97
8	15	.73	.86	.75	.85	.74	.85	.79	.86	.67	.79	.77	.87	.75	.87
	16	.99	NA	1.00	NA	.98	NA	.99	NA	.99	NA	.99	NA	.99	NA
9	17	.86	.81	.82	.84	.85	.81	.82	.85	.73	.83	.86	.80	.86	.78
	18	NA	.61	NA	.67	NA	.64	NA	.63	NA	.59	NA	.66	NA	.66
10	19	.52	.82	.63	.87	.56	.85	.60	.90	.53	.83	.60	.86	.57	.86
	20	.94	.90	.93	.93	.90	.91	.94	.87	.93	.90	.93	.92	.92	.92
11	21	.84	.52	.84	.51	.79	.51	.82	.50	.80	.45	.82	.54	.84	.53
	22	.97	.76	.98	.75	.96	.79	.97	.72	.96	.72	.98	.75	.97	.75
12	23	.84	.96	.87	.98	.82	.94	.92	.97	.84	.97	.87	.98	.85	.98
	24	.91	.85	.94	.87	.91	.85	.88	.80	.89	.80	.92	.85	.91	.85
13	25	.61	.59	.64	.61	.59	.60	.61	.60	.57	.54	.60	.58	.63	.60
	26	.92	.91	.90	.90	.91	.89	.92	.91	.91	.90	.90	.89	.90	.91
14	27	.55	.86	.70	.87	.66	.85	.70	.92	.65	.88	.67	.86	.65	.84
	28	.99	.84	1.00	.89	1.00	.88	.99	.85	.99	.84	.99	.87	.99	.87
15	29	.66	.32	.74	.37	.71	.39	.78	.26	.70	.35	.72	.37	.71	.37
	30	.97	.42	.98	.39	.96	.32	.96	.37	.96	.32	.97	.43	.97	.44
16	31	.80	.78	.81	.81	.82	.81	.87	.85	.83	.81	.82	.82	.83	.80
	32	.80	.80	.85	.83	.86	.80	.81	.68	.79	.73	.83	.82	.82	.83
17	33	.87	.85	.87	.79	.84	.77	.86	.89	.84	.87	.85	.81	.84	.82
	34	.91	.98	.94	.98	.98	.98	.95	.99	.92	.97	.94	.98	.93	.98
18	35	.93	.84	.96	.85	.97	.80	.96	.85	.90	.81	.95	.84	.95	.83
	36	.89	.90	.93	.93	.89	.95	.91	.95	.86	.91	.92	.93	.93	.94
19	37	.60	.41	.69	.49	.67	.49	.74	.56	.61	.45	.66	.47	.64	.45
	38	.59	.96	.52	.97	.60	.96	.68	.98	.58	.96	.55	.97	.55	.96
20	39	.80	.60	.85	.65	.83	.63	.83	.75	.79	.64	.83	.65	.83	.62
	40	.87	.56	.94	.49	.95	.50	.93	.64	.90	.57	.92	.51	.93	.52
21	41	.56	.94	.65	.96	.66	.94	.77	.96	.67	.93	.69	.96	.66	.96
	42	.49	.76	.69	.84	.70	.74	.68	.78	.55	.74	.56	.82	.61	.81
22	43	.94	.90	.97	.91	.98	.91	.95	.90	.93	.87	.96	.92	.96	.90
	44	.76	.93	.84	.93	.77	.96	.78	.94	.73	.92	.82	.95	.81	.93
23	45	.89	.86	.90	.87	.88	.83	.90	.87	.87	.81	.91	.87	.91	.86
	46	.93	.89	.94	.90	.91	.88	.94	.91	.92	.86	.93	.90	.94	.90
24	47	.42	.76	.55	.76	.53	.79	.41	.85	.43	.76	.48	.78	.50	.77
	48	.49	.90	.44	.91	.43	.94	.36	.93	.42	.89	.46	.91	.50	.90
25	49	.87	.66	.87	.64	.83	.61	.87	.71	.82	.65	.88	.67	.88	.67
	50	.89	.97	.91	.97	.91	.97	.92	.99	.85	.97	.91	.97	.91	.98

Note. NA indicates that the parameters could not be estimated, because these items did not have a correct LE response option.

APPENDIX G. Stratified Samples Item-Level Discriminability (ITC)

Item	Response	OTS-CIV		OTS-AD		AECP		USAFA		ROTC		ANG		AFRES	
		T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
1	1	.09	.14	.06	.07	.01	.19	.06	.09	.12	.12	.12	.10	.02	.13
	2	.12	.16	.09	.10	.10	.10	.08	.12	.13	.16	.13	.15	.10	.15
2	3	.10	.19	.07	.14	.09	.09	.07	.15	.09	.15	.09	.15	.08	.20
	4	.13	.18	.12	.11	.06	.16	.08	.14	.14	.15	.14	.13	.09	.16
3	5	.11	.23	.10	.17	.02	.11	.06	.18	.10	.20	.10	.20	.09	.25
	6	.18	.19	.12	.11	.05	.13	.11	.12	.15	.17	.15	.16	.08	.20
4	7	.18	-.03	.13	-.02	.14	-.02	.14	-.04	.15	-.06	.15	.04	.12	.02
	8	.14	.07	.10	.08	.09	.11	.12	.11	.14	.07	.14	.13	.08	.11
5	9	.21	.15	.12	.11	.03	.14	.09	.10	.18	.12	.18	.13	.18	.11
	10	.15	.14	.10	.12	.09	.21	.09	.10	.15	.12	.15	.14	.12	.07
6	11	.15	.05	.08	.05	.11	-.11	.17	.08	.12	.05	.12	.08	.11	.05
	12	.12	.10	.08	.08	.07	.09	.15	.12	.10	.13	.10	.19	-.01	.06
7	13	.10	.21	.11	.16	.01	.19	.09	.17	.10	.21	.10	.16	.12	.17
	14	.11	.18	.10	.12	.10	.21	.09	.11	.11	.15	.11	.08	.12	.11
8	15	.10	.13	.07	.11	.26	.28	.08	.12	.07	.10	.07	.09	.09	.09
	16	.12	.10	.10	.08	.09	.16	.11	.06	.16	.04	.16	.06	.06	.07
9	17	.10	.06	.08	.10	-.07	-.01	.11	.10	.09	.09	.09	.13	.10	.10
	18	.06	.19	.04	.16	-.02	.10	.05	.11	.04	.14	.04	.14	.06	.11
10	19	.09	.11	.09	.09	.03	.08	.07	.08	.10	.09	.10	.10	.06	.06
	20	.10	.17	.05	.16	.02	.18	.14	.15	.11	.15	.11	.19	.03	.18
11	21	.13	.08	.12	.05	.05	.08	.07	.04	.11	.06	.11	.05	.13	.05
	22	.12	.13	.08	.11	.10	.15	.13	.13	.12	.13	.12	.16	.11	.10
12	23	.14	.16	.08	.13	.05	.25	.11	.13	.14	.10	.14	.10	.11	.09
	24	.16	.20	.13	.14	.19	-.02	.17	.14	.16	.14	.16	.16	.17	.17
13	25	.09	.08	.12	.09	.17	.09	.07	.04	.09	.08	.09	.12	.12	.12
	26	.11	.11	.11	.12	.13	.01	.07	.06	.11	.08	.11	.12	.13	.15
14	27	.10	.09	.06	.09	.03	.16	.13	.13	.12	.09	.12	.10	.10	.09
	28	.10	.14	.09	.13	NA	.12	.22	.11	.14	.17	.14	.17	.08	.14
15	29	.12	.11	.09	.08	-.02	.10	.12	.08	.13	.12	.13	.06	.11	.09
	30	.09	.11	.04	.10	.06	.12	.13	.09	.10	.11	.10	.10	.15	.11
16	31	.08	.22	.08	.18	.10	.36	.08	.14	.08	.20	.08	.22	.10	.17
	32	.12	.12	.11	.12	.12	-.02	.09	.06	.13	.10	.13	.17	.08	.14
17	33	.23	.14	.17	.14	.10	.16	.19	.09	.19	.14	.19	.11	.25	.08
	34	.17	.23	.18	.26	-.01	.29	.16	.21	.18	.25	.18	.27	.21	.22
18	35	.18	.29	.15	.27	.07	.37	.13	.21	.20	.26	.20	.23	.20	.21
	36	.23	.26	.19	.28	.15	.36	.20	.21	.24	.24	.24	.27	.23	.26
19	37	.15	.10	.15	.12	.24	.01	.10	.11	.18	.10	.18	.11	.15	.08
	38	.09	.20	.11	.27	-.06	.18	.09	.16	.10	.21	.10	.27	.11	.24
20	39	.10	.16	.12	.17	.10	.17	.12	.17	.12	.17	.12	.17	.17	.15
	40	.17	.12	.22	.14	.07	.10	.17	.09	.20	.11	.20	.13	.30	.13
21	41	.14	.24	.12	.27	.27	.18	.09	.18	.17	.24	.17	.25	.15	.23
	42	.21	.26	.22	.34	.31	.32	.15	.19	.21	.25	.21	.35	.23	.32
22	43	.21	.29	.20	.34	.00	.13	.19	.19	.22	.23	.22	.36	.25	.30
	44	.21	.25	.26	.30	.18	.08	.14	.21	.19	.25	.19	.27	.29	.29
23	45	.25	.32	.26	.37	.20	.31	.16	.26	.22	.31	.22	.32	.25	.33
	46	.22	.30	.29	.32	.17	.34	.18	.19	.22	.26	.22	.33	.25	.25
24	47	-.08	.31	-.03	.31	.08	.10	-.05	.23	-.04	.30	-.04	.33	-.04	.30
	48	.02	.27	.02	.30	-.07	.13	.00	.20	-.02	.26	-.02	.29	.04	.29
25	49	.22	.25	.26	.21	.19	.33	.18	.15	.24	.23	.24	.26	.24	.28
	50	.22	.17	.23	.18	.14	-.06	.17	.13	.22	.17	.22	.18	.28	.14

Note. NA indicates that the parameters could not be estimated due to lack of variability.

APPENDIX H. Stratified Samples Item-Level Internal Consistency if Item is Deleted (Cronbach’s alpha)

Item	Response	OTS-CIV		OTS-AD		AECP		USAFA		ROTC		ANG		AFRES	
		T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
1	1	.56	.64	.52	.63	.45	.58	.48	.54	.56	.61	.50	.65	.56	.62
	2	.56	.64	.52	.63	.45	.59	.48	.54	.56	.61	.51	.65	.56	.62
2	3	.56	.64	.52	.62	.45	.59	.48	.54	.56	.61	.50	.65	.56	.62
	4	.56	.64	.52	.63	.45	.58	.48	.54	.56	.61	.50	.65	.56	.62
3	5	.56	.64	.52	.62	.45	.59	.48	.54	.56	.61	.51	.65	.56	.62
	6	.55	.64	.52	.63	.45	.58	.48	.54	.56	.61	.50	.65	.56	.62
4	7	.55	.66	.51	.64	.44	.60	.47	.57	.55	.64	.50	.66	.56	.64
	8	.56	.65	.52	.63	.45	.58	.48	.54	.56	.62	.51	.65	.56	.63
5	9	.55	.64	.52	.63	.45	.58	.48	.54	.55	.61	.50	.65	.56	.63
	10	.55	.64	.52	.63	.45	.58	.48	.54	.56	.61	.50	.65	.56	.63
6	11	.55	.66	.52	.64	.45	.61	.47	.55	.56	.62	.50	.66	.56	.64
	12	.56	.65	.52	.63	.45	.59	.48	.55	.56	.61	.51	.65	.56	.63
7	13	.56	.64	.52	.62	.46	.58	.48	.54	.56	.61	.50	.65	.56	.62
	14	.56	.64	.52	.63	.45	.58	.49	.54	.56	.61	.50	.65	.56	.63
8	15	.56	.64	.52	.63	.42	.58	.48	.54	.57	.61	.51	.65	.56	.63
	16	.56	.65	.52	.63	.45	.58	.48	.54	.56	.62	.51	.65	.56	.63
9	17	.56	.65	.52	.63	.46	.59	.48	.54	.56	.62	.51	.65	.56	.63
	18	.56	.64	.52	.62	.46	.59	.48	.54	.56	.61	.51	.65	.56	.63
10	19	.56	.65	.52	.63	.46	.59	.49	.54	.56	.62	.51	.65	.57	.63
	20	.56	.64	.52	.62	.45	.58	.48	.54	.56	.61	.50	.65	.56	.62
11	21	.55	.65	.51	.63	.45	.59	.49	.55	.56	.62	.51	.65	.56	.63
	22	.56	.64	.52	.63	.45	.58	.48	.54	.56	.61	.50	.65	.56	.63
12	23	.56	.64	.52	.63	.45	.58	.48	.54	.56	.61	.50	.65	.56	.64
	24	.55	.64	.51	.62	.44	.59	.47	.54	.56	.61	.50	.65	.55	.62
13	25	.56	.65	.51	.64	.43	.59	.49	.55	.56	.62	.50	.65	.56	.63
	26	.56	.64	.52	.63	.44	.59	.48	.54	.56	.62	.50	.65	.56	.62
14	27	.56	.65	.52	.63	.46	.58	.48	.54	.56	.62	.51	.65	.56	.63
	28	.56	.65	.52	.62	.45	.58	.48	.54	.56	.61	.51	.65	.56	.62
15	29	.56	.65	.52	.63	.47	.59	.48	.54	.56	.61	.51	.65	.56	.63
	30	.56	.65	.52	.63	.45	.58	.48	.54	.56	.61	.50	.65	.56	.63
16	31	.56	.64	.52	.62	.45	.57	.48	.54	.56	.61	.51	.64	.56	.62
	32	.55	.64	.52	.63	.44	.59	.48	.54	.56	.61	.50	.65	.56	.62
17	33	.55	.64	.51	.63	.45	.58	.47	.54	.55	.61	.50	.65	.55	.63
	34	.55	.64	.51	.62	.45	.58	.48	.54	.56	.61	.50	.65	.55	.62
18	35	.55	.64	.51	.62	.45	.56	.48	.53	.55	.61	.50	.65	.55	.62
	36	.55	.64	.51	.62	.44	.58	.47	.54	.55	.61	.50	.65	.55	.62
19	37	.55	.65	.51	.63	.44	.60	.48	.54	.55	.62	.49	.65	.56	.63
	38	.56	.64	.51	.62	.47	.58	.48	.54	.56	.61	.50	.65	.56	.62
20	39	.56	.64	.51	.62	.45	.58	.48	.53	.56	.61	.50	.65	.55	.62
	40	.55	.65	.51	.62	.45	.59	.47	.54	.55	.62	.49	.65	.55	.63
21	41	.55	.64	.52	.62	.43	.58	.48	.54	.56	.61	.50	.65	.56	.62
	42	.55	.63	.50	.61	.42	.56	.47	.53	.55	.60	.49	.63	.55	.61
22	43	.55	.63	.51	.61	.45	.58	.48	.53	.56	.61	.50	.64	.56	.61
	44	.54	.64	.50	.61	.43	.59	.47	.53	.55	.61	.49	.64	.54	.62
23	45	.54	.63	.50	.61	.43	.57	.47	.53	.55	.60	.49	.64	.55	.61
	46	.55	.64	.50	.61	.44	.57	.47	.53	.55	.60	.49	.64	.55	.62
24	47	.58	.63	.54	.61	.45	.59	.51	.53	.58	.60	.53	.63	.58	.61
	48	.57	.64	.53	.62	.48	.58	.50	.53	.58	.61	.52	.64	.58	.62
25	49	.55	.63	.50	.62	.44	.56	.47	.54	.55	.60	.49	.64	.55	.61
	50	.55	.64	.51	.62	.44	.59	.47	.54	.55	.61	.49	.65	.55	.62

APPENDIX I. Stratified Samples Item-Level Subgroup Differences (Cohen’s d)

OTS-CIV

Item	Response	T1 Cohen’s *d**					T2 Cohen’s *d**				
		F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
1	1	.09	.01	.06	.02	.05	.01	.21	.01	.02	.06
	2	.02	.09	.15	.11	.10	.02	.08	.05	.05	.03
2	3	.19	.12	.08	.09	.00	.03	.22	.08	.11	.05
	4	.02	.20	.10	.09	.09	.03	.05	.07	.05	.00
3	5	.00	.07	.04	.03	.03	.04	.14	.19	.07	.02
	6	.01	.05	.03	.01	.04	.02	.11	.17	.05	.02
4	7	.03	.21	.08	.08	.05	.06	.02	.03	.04	.17
	8	.03	.10	.01	.01	.17	.06	.25	.16	.16	.09
5	9	.02	.15	.18	.05	.00	.27	.13	.10	.06	.10
	10	.03	.12	.17	.05	.04	.11	.09	.12	.15	.12
6	11	.03	.13	.04	.01	.02	.02	.06	.09	.01	.12
	12	.04	.05	.09	.04	.05	.05	.06	.07	.01	.13
7	13	.08	.13	.18	.10	.10	.05	.17	.27	.10	.26
	14	.13	.13	.15	.12	.18	.01	.13	.16	.11	.38
8	15	.05	.10	.04	.06	.14	.02	.08	.26	.11	.24
	16	.02	.08	.07	.11	.07	.02	.12	.16	.02	.16
9	17	.04	.10	.16	.06	.12	.01	.02	.16	.00	.09
	18	.01	.06	.24	.01	.08	.08	.19	.37	.17	.14
10	19	.14	.06	.04	.02	.14	.12	.22	.18	.09	.03
	20	.02	.02	.04	.03	.03	.07	.07	.06	.03	.15
11	21	.01	.26	.20	.04	.03	.04	.24	.11	.07	.03
	22	.05	.14	.21	.03	.10	.01	.13	.14	.01	.04
12	23	.07	.28	.19	.16	.04	.05	.04	.06	.03	.08
	24	.02	.11	.04	.04	.09	.06	.12	.13	.02	.00
13	25	.00	.13	.02	.01	.06	.03	.14	.03	.07	.10
	26	.00	.10	.16	.06	.08	.02	.15	.08	.03	.22
14	27	.17	.10	.06	.07	.12	.03	.05	.15	.17	.10
	28	.01	.08	.04	.04	.13	.12	.14	.14	.13	.09
15	29	.07	.21	.11	.08	.10	.03	.10	.01	.11	.04
	30	.02	.08	.04	.01	.06	.15	.08	.14	.05	.33
16	31	.04	.10	.05	.05	.13	.04	.20	.00	.13	.09
	32	.14	.11	.17	.01	.02	.09	.01	.11	.00	.05
17	33	.03	.22	.25	.05	.00	.03	.30	.16	.11	.05
	34	.08	.17	.14	.01	.04	.01	.18	.07	.04	.12
18	35	.08	.32	.33	.14	.25	.05	.25	.29	.05	.07
	36	.09	.28	.22	.04	.24	.06	.20	.18	.14	.02
19	37	.07	.41	.25	.11	.07	.14	.04	.01	.07	.08
	38	.10	.22	.02	.04	.16	.02	.09	.01	.08	.01
20	39	.11	.12	.11	.04	.08	.11	.39	.23	.10	.01
	40	.04	.23	.14	.10	.16	.08	.24	.12	.03	.29
21	41	.02	.27	.12	.12	.05	.03	.27	.11	.08	.09
	42	.15	.31	.24	.18	.04	.09	.31	.36	.12	.09
22	43	.03	.26	.12	.06	.12	.05	.35	.45	.11	.04
	44	.08	.32	.31	.06	.00	.05	.28	.32	.06	.16
23	45	.06	.30	.42	.07	.06	.01	.40	.23	.21	.13
	46	.00	.25	.31	.01	.02	.01	.33	.19	.15	.05
24	47	.04	.15	.11	.05	.01	.11	.37	.23	.15	.07
	48	.01	.14	.10	.10	.20	.07	.18	.09	.11	.03
25	49	.02	.34	.19	.08	.01	.07	.34	.22	.17	.10
	50	.01	.33	.10	.04	.01	.04	.32	.10	.14	.08

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic. N_{T1} = 8,304; N_{T2} = 7,601.

OTS-AD

Item	Response	T1 Cohen’s d*					T2 Cohen’s d*				
		F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
1	1	.07	.04	.08	.03	.04	.04	.19	.09	.03	.08
	2	.02	.10	.09	.03	.08	.01	.02	.07	.01	.10
2	3	.23	.10	.05	.05	.03	.07	.20	.05	.02	.15
	4	.00	.18	.05	.01	.01	.00	.11	.11	.02	.12
3	5	.01	.13	.06	.07	.08	.11	.17	.13	.07	.02
	6	.06	.02	.03	.01	.08	.02	.18	.09	.04	.02
4	7	.08	.15	.07	.02	.02	.02	.10	.00	.02	.16
	8	.06	.07	.09	.04	.03	.00	.15	.21	.07	.01
5	9	.02	.17	.12	.05	.10	.24	.17	.09	.05	.15
	10	.02	.12	.17	.05	.02	.05	.15	.05	.07	.19
6	11	.01	.09	.12	.04	.23	.04	.06	.09	.04	.09
	12	.02	.01	.09	.03	.08	.00	.06	.18	.13	.11
7	13	.06	.12	.11	.00	.03	.05	.26	.24	.04	.20
	14	.10	.11	.09	.02	.12	.01	.23	.05	.02	.05
8	15	.08	.04	.05	.10	.02	.00	.13	.16	.12	.02
	16	.05	.09	.11	.04	.00	.07	.02	.06	.08	.00
9	17	.03	.12	.26	.05	.18	.08	.04	.11	.06	.05
	18	.02	.11	.21	.06	.18	.13	.12	.38	.22	.05
10	19	.13	.09	.04	.00	.04	.16	.23	.19	.04	.03
	20	.01	.01	.13	.00	.06	.04	.19	.02	.02	.03
11	21	.01	.29	.21	.14	.17	.06	.10	.07	.02	.17
	22	.02	.09	.17	.03	.04	.02	.03	.09	.01	.16
12	23	.15	.27	.11	.01	.10	.01	.03	.00	.01	.17
	24	.01	.08	.09	.03	.08	.03	.10	.05	.07	.02
13	25	.03	.18	.08	.02	.02	.03	.18	.00	.06	.02
	26	.04	.11	.20	.11	.11	.01	.14	.04	.13	.23
14	27	.06	.11	.17	.07	.14	.05	.08	.16	.05	.05
	28	.03	.09	.02	.06	.10	.05	.09	.14	.16	.11
15	29	.15	.19	.08	.06	.06	.05	.08	.00	.04	.07
	30	.06	.06	.01	.04	.10	.10	.04	.15	.07	.16
16	31	.01	.08	.08	.00	.08	.00	.16	.09	.09	.12
	32	.18	.03	.16	.08	.08	.01	.10	.21	.04	.01
17	33	.02	.24	.32	.12	.01	.05	.21	.11	.13	.05
	34	.05	.22	.23	.11	.03	.04	.14	.05	.01	.09
18	35	.08	.24	.34	.15	.15	.05	.33	.26	.14	.02
	36	.04	.25	.32	.13	.08	.06	.24	.30	.16	.04
19	37	.11	.33	.12	.06	.14	.09	.07	.03	.02	.08
	38	.10	.23	.21	.09	.05	.00	.20	.26	.10	.10
20	39	.11	.05	.24	.05	.15	.11	.45	.20	.21	.03
	40	.01	.17	.10	.03	.01	.02	.22	.09	.20	.10
21	41	.02	.23	.02	.09	.08	.01	.17	.02	.08	.13
	42	.00	.30	.13	.08	.16	.00	.42	.31	.10	.07
22	43	.01	.22	.14	.13	.01	.07	.33	.34	.15	.06
	44	.06	.47	.36	.11	.06	.04	.38	.25	.05	.07
23	45	.01	.36	.35	.16	.13	.00	.39	.19	.10	.04
	46	.04	.30	.23	.08	.02	.06	.37	.31	.13	.07
24	47	.02	.16	.11	.14	.03	.05	.39	.21	.25	.04
	48	.04	.06	.17	.11	.11	.08	.30	.12	.19	.00
25	49	.02	.27	.23	.14	.04	.06	.28	.26	.04	.14
	50	.08	.38	.34	.14	.11	.05	.18	.10	.15	.15

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.
Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic. N_{T1} = 5,890; N_{T2} = 5,275.

Item	Response	T1 Cohen's *d**					T2 Cohen's *d**				
		F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
1	1	.28	.13	.53	.04	NA	.03	.33	.50	.04	NA
	2	.10	.13	.14	.16	NA	.14	.23	.21	.24	NA
2	3	.16	.01	.35	.38	NA	.06	.12	.20	.24	NA
	4	.00	.03	.19	.22	NA	.10	.13	.23	.70	NA
3	5	.02	.31	.10	.06	NA	.07	.34	.54	.28	NA
	6	.12	.06	.17	.44	NA	.11	.07	.72	.70	NA
4	7	.06	.28	.10	.00	NA	.08	.19	.04	.34	NA
	8	.24	.12	.59	.26	NA	.06	.20	.45	.23	NA
5	9	.22	.12	.11	.01	NA	.36	.20	.60	.14	NA
	10	.08	.06	.27	.13	NA	.06	.04	.89	.21	NA
6	11	.18	.22	.38	.26	NA	.36	.08	.06	.41	NA
	12	.32	.33	.29	.06	NA	.14	.15	.10	.15	NA
7	13	.03	.39	.76	.25	NA	.00	.19	.23	.23	NA
	14	.05	.12	.05	.13	NA	.25	.25	.21	.30	NA
8	15	.21	.22	.03	.33	NA	.60	.35	.38	.13	NA
	16	.08	.11	1.06	.11	NA	.27	.39	.39	.03	NA
9	17	.09	.26	.08	.25	NA	.11	.25	.16	.57	NA
	18	.03	.08	.12	.23	NA	.07	.18	.65	.25	NA
10	19	.10	.19	.24	.58	NA	.14	.36	.14	.50	NA
	20	.02	.08	.55	.17	NA	.35	.03	.27	.31	NA
11	21	.10	.28	.21	.12	NA	.04	.15	.06	.02	NA
	22	.24	.23	.10	.11	NA	.12	.14	.74	.18	NA
12	23	.01	.26	.46	.08	NA	.01	.15	.21	.31	NA
	24	.14	.19	.34	.13	NA	.33	.14	.03	.43	NA
13	25	.09	.28	.15	.24	NA	.04	.26	.06	.00	NA
	26	.05	.42	.44	.18	NA	.01	.59	.34	.06	NA
14	27	.01	.06	.51	.14	NA	.02	.09	.08	.45	NA
	28	NA	NA	NA	NA	NA	.12	.50	.01	.19	NA
15	29	.02	.24	.70	.19	NA	.03	.45	.39	.19	NA
	30	.06	.25	.17	.19	NA	.05	.18	.06	.42	NA
16	31	.11	.58	.44	.04	NA	.64	.51	.08	.57	NA
	32	.09	.47	.33	.17	NA	.26	.46	.46	.18	NA
17	33	.15	.22	.38	.10	NA	.04	.19	.85	.62	NA
	34	.15	.18	.16	.08	NA	.30	NA	1.39	NA	NA
18	35	.06	.00	.22	.04	NA	.08	.48	.01	.09	NA
	36	.06	.41	.14	.39	NA	.13	.17	1.64	.63	NA
19	37	.40	.54	.43	.12	NA	.17	.28	.47	.12	NA
	38	.25	.54	.29	.18	NA	.04	.18	.16	.72	NA
20	39	.11	.20	.23	.08	NA	.24	.43	.23	.21	NA
	40	.15	.54	.17	.58	NA	.07	.49	.35	.41	NA
21	41	.13	.30	.40	.00	NA	.22	.04	.20	.07	NA
	42	.07	.21	.47	.25	NA	.09	.57	.12	.72	NA
22	43	.19	.11	.10	.14	NA	.12	.17	.16	.30	NA
	44	.26	.00	.66	.14	NA	.06	.12	.94	.45	NA
23	45	.04	.13	.13	.25	NA	.39	.37	.97	.73	NA
	46	.09	.10	.21	.02	NA	.37	.45	.24	.09	NA
24	47	.17	.30	.09	.23	NA	.19	.19	.11	.32	NA
	48	.15	.16	.25	.40	NA	.09	.35	.20	.79	NA
25	49	.39	.05	.63	.17	NA	.25	.58	.35	.05	NA
	50	.29	.66	.26	.15	NA	.04	.11	.10	.15	NA

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic. NA indicates that there was little to no variability on some items and the sample size was too small to calculate Cohen's *d*. N_{T1} = 175; N_{T2} = 172.

USAFA

Item	Response	T1 Cohen's *d**					T2 Cohen's *d**				
		F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
1	1	.12	.03	.07	.08	.01	.07	.12	.05	.02	.26
	2	.06	.05	.02	.05	.07	.02	.05	.09	.04	.04
2	3	.02	.05	.16	.11	.05	.00	.19	.03	.13	.09
	4	.00	.14	.06	.04	.21	.02	.08	.09	.00	.17
3	5	.04	.14	.09	.03	.08	.01	.08	.03	.02	.21
	6	.06	.10	.04	.01	.18	.03	.10	.12	.01	.10
4	7	.01	.33	.01	.03	.14	.01	.20	.04	.11	.58
	8	.02	.11	.08	.01	.31	.03	.17	.15	.07	.01
5	9	.04	.05	.17	.03	.14	.11	.12	.02	.05	.34
	10	.01	.02	.16	.05	.17	.03	.08	.01	.07	.03
6	11	.14	.12	.02	.09	.10	.05	.02	.05	.07	.15
	12	.07	.08	.05	.07	.20	.03	.03	.01	.06	.09
7	13	.11	.15	.18	.00	.08	.06	.10	.10	.04	.14
	14	.22	.12	.14	.02	.07	.01	.15	.04	.04	.24
8	15	.02	.08	.10	.08	.05	.03	.23	.09	.02	.19
	16	.03	.03	.07	.05	.10	.01	.09	.17	.06	.08
9	17	.04	.21	.24	.07	.02	.10	.06	.02	.08	.01
	18	.01	.12	.31	.08	.05	.07	.09	.29	.02	.19
10	19	.10	.14	.23	.03	.20	.01	.04	.11	.02	.04
	20	.04	.02	.10	.04	.05	.02	.23	.08	.05	.10
11	21	.04	.19	.06	.03	.03	.06	.03	.05	.01	.02
	22	.08	.26	.15	.07	.09	.02	.16	.20	.07	.19
12	23	.01	.10	.08	.08	.16	.05	.02	.05	.01	.15
	24	.09	.13	.07	.02	.19	.12	.14	.06	.14	.30
13	25	.00	.08	.14	.01	.21	.01	.05	.13	.03	.23
	26	.00	.03	.03	.04	.01	.02	.03	.09	.01	.33
14	27	.09	.21	.13	.02	.10	.06	.09	.19	.02	.13
	28	.05	.15	.09	.03	.04	.13	.15	.07	.07	.19
15	29	.00	.10	.08	.03	.02	.03	.01	.08	.07	.15
	30	.01	.04	.10	.02	.11	.03	.04	.06	.03	.15
16	31	.04	.08	.09	.00	.04	.00	.04	.02	.09	.10
	32	.15	.05	.07	.03	.15	.16	.14	.05	.06	.33
17	33	.07	.18	.16	.03	.08	.01	.16	.03	.23	.10
	34	.02	.12	.15	.06	.07	.03	.23	.05	.01	.04
18	35	.03	.22	.15	.03	.12	.01	.28	.18	.04	.14
	36	.06	.21	.13	.03	.06	.05	.15	.10	.06	.09
19	37	.00	.21	.07	.01	.12	.08	.02	.07	.05	.15
	38	.01	.20	.03	.06	.02	.02	.11	.05	.01	.19
20	39	.15	.21	.12	.01	.11	.01	.04	.00	.01	.46
	40	.01	.19	.01	.02	.28	.00	.19	.00	.01	.09
21	41	.11	.32	.08	.06	.18	.08	.17	.04	.06	.15
	42	.05	.35	.27	.07	.11	.09	.24	.13	.10	.04
22	43	.02	.29	.04	.08	.07	.06	.17	.12	.20	.27
	44	.12	.27	.11	.12	.02	.02	.11	.16	.12	.03
23	45	.00	.35	.30	.06	.18	.02	.18	.10	.01	.23
	46	.01	.24	.10	.01	.13	.01	.17	.10	.04	.18
24	47	.01	.18	.09	.11	.00	.13	.19	.02	.09	.08
	48	.07	.18	.08	.00	.16	.03	.10	.04	.06	.07
25	49	.03	.24	.11	.02	.05	.11	.15	.20	.02	.06
	50	.04	.24	.17	.05	.06	.09	.14	.08	.00	.16

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic. N_{T1} = 4,611; N_{T2} = 3,962.

ROTC

Item	Response	T1 Cohen's *d**					T2 Cohen's *d**				
		F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
1	1	.06	.02	.02	.01	.07	.00	.02	.03	.07	.08
	2	.03	.09	.11	.07	.01	.05	.17	.01	.10	.02
2	3	.02	.10	.12	.06	.15	.01	.25	.01	.07	.08
	4	.01	.18	.08	.13	.07	.02	.15	.02	.05	.04
3	5	.04	.01	.06	.05	.10	.02	.22	.16	.08	.12
	6	.01	.05	.05	.08	.03	.03	.17	.16	.09	.16
4	7	.01	.26	.04	.11	.06	.03	.10	.00	.03	.05
	8	.01	.16	.05	.03	.10	.02	.20	.06	.04	.13
5	9	.04	.17	.12	.04	.10	.22	.19	.03	.10	.02
	10	.02	.14	.17	.00	.16	.01	.21	.04	.05	.00
6	11	.01	.12	.02	.01	.00	.02	.11	.06	.02	.03
	12	.02	.01	.01	.00	.00	.02	.13	.00	.08	.01
7	13	.11	.13	.17	.10	.02	.03	.31	.24	.12	.09
	14	.18	.14	.08	.04	.07	.01	.17	.07	.08	.01
8	15	.02	.13	.13	.05	.08	.05	.10	.16	.08	.08
	16	.01	.10	.03	.10	.03	.01	.03	.14	.04	.13
9	17	.03	.06	.10	.05	.08	.03	.04	.05	.01	.10
	18	.02	.02	.18	.12	.04	.12	.19	.28	.16	.06
10	19	.09	.08	.03	.07	.15	.13	.22	.14	.07	.23
	20	.00	.05	.01	.05	.09	.02	.09	.02	.03	.03
11	21	.01	.30	.18	.11	.05	.01	.11	.05	.05	.08
	22	.02	.14	.15	.10	.05	.03	.03	.15	.03	.08
12	23	.13	.30	.16	.09	.10	.04	.04	.04	.02	.06
	24	.02	.11	.03	.07	.09	.05	.07	.00	.06	.05
13	25	.01	.15	.03	.01	.03	.00	.14	.03	.01	.02
	26	.01	.12	.06	.01	.10	.03	.12	.02	.01	.09
14	27	.06	.18	.07	.08	.14	.05	.08	.08	.03	.05
	28	.01	.09	.03	.02	.11	.11	.14	.08	.11	.06
15	29	.04	.23	.14	.04	.05	.04	.08	.13	.05	.11
	30	.01	.09	.05	.04	.07	.10	.04	.08	.04	.04
16	31	.03	.10	.13	.00	.09	.03	.30	.06	.13	.06
	32	.13	.18	.06	.00	.02	.11	.01	.03	.00	.00
17	33	.02	.20	.12	.08	.02	.06	.25	.06	.09	.05
	34	.02	.16	.14	.12	.00	.04	.32	.05	.11	.07
18	35	.14	.34	.23	.12	.04	.01	.32	.17	.11	.10
	36	.13	.28	.14	.12	.02	.03	.35	.07	.16	.05
19	37	.11	.49	.18	.10	.15	.17	.07	.05	.06	.03
	38	.12	.20	.05	.02	.01	.05	.23	.03	.16	.06
20	39	.11	.12	.20	.09	.10	.09	.44	.17	.14	.02
	40	.03	.25	.12	.17	.13	.08	.23	.11	.08	.05
21	41	.10	.28	.14	.10	.02	.01	.33	.12	.20	.14
	42	.12	.36	.20	.14	.15	.01	.37	.20	.13	.12
22	43	.03	.30	.10	.09	.12	.04	.28	.21	.13	.03
	44	.01	.31	.21	.10	.07	.02	.27	.16	.15	.01
23	45	.06	.26	.24	.19	.19	.07	.47	.18	.19	.05
	46	.02	.23	.17	.14	.06	.11	.42	.19	.19	.11
24	47	.06	.12	.12	.08	.04	.05	.45	.14	.23	.01
	48	.01	.22	.14	.12	.03	.03	.26	.01	.15	.08
25	49	.04	.40	.11	.19	.07	.03	.24	.08	.08	.10
	50	.08	.33	.12	.19	.01	.01	.16	.05	.08	.01

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic. N_{T1} = 14,591; N_{T2} = 14,392.

ANG

Item	Response	T1 Cohen's *d**					T2 Cohen's *d**				
		F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
1	1	.07	.03	.02	.01	.00	.00	.12	.04	.00	.24
	2	.05	.06	.08	.05	.17	.04	.05	.06	.00	.03
2	3	.13	.17	.13	.03	.10	.00	.17	.14	.04	.05
	4	.05	.20	.21	.05	.09	.10	.05	.11	.02	.05
3	5	.04	.10	.03	.10	.03	.02	.11	.17	.08	.03
	6	.00	.08	.14	.00	.05	.05	.11	.11	.16	.02
4	7	.00	.12	.01	.04	.01	.09	.06	.02	.05	.19
	8	.00	.04	.04	.01	.25	.01	.32	.15	.05	.17
5	9	.04	.14	.07	.03	.27	.28	.21	.03	.14	.25
	10	.06	.13	.31	.08	.24	.01	.18	.25	.11	.03
6	11	.01	.09	.09	.03	.02	.03	.12	.08	.05	.24
	12	.01	.10	.12	.06	.34	.06	.08	.01	.01	.13
7	13	.04	.13	.09	.01	.05	.12	.22	.08	.01	.03
	14	.10	.18	.03	.07	.11	.02	.01	.04	.13	.14
8	15	.03	.09	.18	.11	.01	.02	.16	.11	.02	.25
	16	.05	.07	.07	.08	.09	.01	.09	.24	.04	.28
9	17	.03	.03	.04	.01	.07	.08	.04	.04	.03	.07
	18	.04	.00	.23	.11	.40	.13	.06	.33	.18	.01
10	19	.15	.08	.08	.07	.28	.03	.35	.04	.02	.20
	20	.02	.03	.13	.08	.13	.06	.03	.15	.20	.25
11	21	.04	.26	.36	.05	.14	.06	.18	.00	.03	.47
	22	.10	.10	.17	.04	.11	.07	.06	.07	.00	.19
12	23	.02	.24	.06	.04	.05	.01	.03	.00	.03	.12
	24	.07	.07	.03	.06	.05	.01	.14	.13	.02	.01
13	25	.07	.20	.06	.07	.10	.00	.11	.15	.05	.05
	26	.07	.07	.08	.00	.13	.03	.17	.04	.01	.08
14	27	.08	.03	.01	.01	.47	.02	.12	.19	.03	.18
	28	.04	.03	.00	.01	.08	.12	.09	.18	.08	.17
15	29	.07	.26	.10	.07	.01	.05	.05	.08	.04	.13
	30	.08	.06	.20	.12	.19	.13	.03	.03	.01	.47
16	31	.04	.12	.17	.11	.15	.02	.29	.02	.04	.23
	32	.19	.06	.18	.09	.13	.08	.10	.14	.07	.12
17	33	.10	.25	.36	.06	.05	.07	.19	.18	.02	.05
	34	.08	.23	.23	.00	.08	.02	.19	.11	.00	.05
18	35	.02	.23	.29	.10	.39	.08	.04	.16	.03	.05
	36	.06	.23	.18	.07	.21	.07	.22	.14	.04	.11
19	37	.10	.44	.24	.11	.13	.06	.13	.05	.09	.35
	38	.08	.22	.15	.05	.21	.01	.20	.05	.00	.30
20	39	.14	.03	.17	.02	.13	.01	.34	.08	.05	.26
	40	.08	.21	.21	.08	.34	.05	.24	.01	.06	.20
21	41	.06	.30	.12	.12	.37	.06	.01	.18	.05	.07
	42	.01	.24	.47	.10	.09	.02	.37	.40	.19	.33
22	43	.02	.25	.27	.02	.11	.02	.23	.00	.07	.21
	44	.04	.33	.32	.11	.05	.05	.31	.11	.07	.26
23	45	.07	.35	.38	.13	.16	.01	.37	.31	.17	.26
	46	.06	.14	.15	.14	.10	.07	.35	.13	.17	.17
24	47	.00	.04	.12	.11	.19	.06	.32	.07	.09	.16
	48	.10	.10	.16	.11	.04	.03	.39	.03	.04	.27
25	49	.04	.15	.23	.09	.41	.05	.23	.01	.13	.28
	50	.00	.25	.20	.15	.26	.04	.15	.14	.04	.04

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.
Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic. N_{T1} = 4,190; N_{T2} = 3,604.

AFRES

Item	Response	T1 Cohen's *d**					T2 Cohen's *d**				
		F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
1	1	.08	.11	.16	.08	.15	.04	.15	.15	.08	.16
	2	.02	.15	.09	.05	.08	.04	.05	.12	.02	.25
2	3	.13	.02	.30	.05	.43	.13	.17	.08	.00	.37
	4	.04	.14	.01	.04	.02	.01	.15	.07	.03	.04
3	5	.07	.11	.12	.06	.29	.06	.06	.07	.18	.14
	6	.06	.01	.08	.08	.06	.01	.22	.12	.12	.05
4	7	.01	.17	.04	.04	.07	.12	.02	.14	.05	.11
	8	.05	.10	.08	.18	.22	.02	.37	.03	.00	.09
5	9	.08	.31	.16	.02	.14	.25	.32	.08	.00	.04
	10	.04	.22	.26	.21	.02	.04	.08	.06	.01	.12
6	11	.02	.22	.20	.04	.22	.04	.14	.02	.05	.27
	12	.12	.13	.07	.16	.11	.04	.06	.05	.05	.10
7	13	.12	.17	.06	.04	.17	.05	.11	.31	.07	.43
	14	.20	.20	.13	.09	.08	.03	.07	.02	.05	.26
8	15	.02	.11	.29	.14	.14	.06	.03	.30	.01	.18
	16	.01	.04	.09	.27	.25	.02	.04	.61	.07	.24
9	17	.06	.27	.24	.04	.07	.08	.11	.02	.04	.27
	18	.01	.10	.26	.09	.13	.16	.27	.26	.11	.00
10	19	.26	.16	.18	.07	.13	.07	.33	.11	.17	.02
	20	.04	.05	.20	.10	.09	.01	.03	.01	.11	.01
11	21	.10	.15	.30	.02	.12	.08	.11	.12	.12	.03
	22	.12	.09	.24	.16	.02	.06	.09	.35	.06	.30
12	23	.04	.33	.02	.15	.23	.04	.35	.14	.00	.12
	24	.11	.22	.24	.15	.09	.03	.11	.03	.06	.25
13	25	.04	.06	.06	.09	.11	.06	.33	.02	.06	.03
	26	.14	.08	.02	.07	.06	.01	.38	.24	.02	.25
14	27	.10	.02	.21	.15	.12	.04	.01	.00	.22	.27
	28	.04	.18	.08	.03	.30	.12	.04	.05	.00	.06
15	29	.14	.29	.13	.04	.39	.05	.26	.05	.12	.37
	30	.04	.18	.11	.07	.02	.08	.09	.01	.04	.02
16	31	.01	.12	.27	.14	.08	.06	.30	.26	.10	.03
	32	.19	.24	.14	.11	.02	.05	.08	.00	.03	.24
17	33	.01	.38	.09	.22	.30	.03	.13	.40	.14	.17
	34	.08	.28	.17	.11	.00	.06	.42	.30	.05	.14
18	35	.06	.33	.12	.13	.14	.08	.14	.10	.05	.25
	36	.04	.41	.31	.08	.09	.03	.46	.05	.02	.40
19	37	.11	.46	.11	.06	.15	.01	.22	.18	.03	.00
	38	.09	.18	.08	.12	.03	.01	.25	.03	.08	.14
20	39	.09	.00	.09	.02	.17	.22	.41	.42	.14	.02
	40	.02	.32	.24	.06	.28	.02	.31	.20	.08	.07
21	41	.02	.37	.03	.02	.03	.08	.21	.02	.06	.28
	42	.01	.34	.17	.24	.31	.09	.36	.44	.02	.37
22	43	.09	.30	.30	.17	.28	.06	.35	.21	.08	.14
	44	.03	.30	.36	.19	.02	.05	.41	.04	.19	.11
23	45	.00	.37	.30	.10	.08	.01	.44	.20	.23	.06
	46	.01	.42	.03	.15	.44	.01	.31	.23	.09	.11
24	47	.06	.16	.20	.04	.35	.05	.28	.10	.05	.11
	48	.09	.00	.10	.12	.01	.13	.27	.02	.03	.27
25	49	.04	.47	.09	.04	.01	.01	.44	.28	.28	.05
	50	.15	.45	.16	.04	.19	.09	.14	.06	.04	.23

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.
Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic. N_{T1} = 1,584; N_{T2} = 1,358.

APPENDIX J. Stratified Samples Common Items Analysis

Item		Response		OTS-CIV			OTS-AD			AECP			USAFA			ROTC			ANG			AFRES		
				p-value		*p*-value	*p*-value		*p*-value	*p*-value		*p*-value	*p*-value		*p*-value	*p*-value		*p*-value	*p*-value		*p*-value	*p*-value		*p*-value
T1	T2	T1	T2	T1	T2	difference	T1	T2	difference	T1	T2	difference	T1	T2	difference	T1	T2	difference	T1	T2	difference	T1	T2	difference
3	1	5	1	.83	.83	.00	.80	.82	.02	.79	.80	.01	.76	.78	.02	.77	.79	.02	.78	.82	.03	.77	.81	.04
		6	2	.92	.92	.00	.94	.94	.00	.95	.95	.00	.96	.96	.00	.95	.95	.00	.94	.95	.01	.95	.94	.01
4	2	7	3	.47	.45	.02	.52	.52	.00	.48	.42	.06	.53	.51	.02	.42	.39	.03	.54	.52	.02	.53	.53	.00
		8	4	.95	.94	.01	.98	.96	.01	.95	.96	.01	.95	.94	.01	.94	.93	.00	.98	.96	.01	.96	.96	.01
5	3	9	5	.84	.83	.01	.90	.89	.01	.89	.89	.01	.92	.91	.01	.76	.75	.01	.88	.87	.01	.89	.88	.01
		10	6	.50	.47	.03	.56	.57	.01	.54	.57	.04	.60	.58	.02	.45	.43	.02	.61	.60	.01	.59	.57	.02
8	6	15	11	.73	.69	.04	.75	.71	.04	.74	.66	.08	.79	.77	.03	.67	.61	.06	.77	.73	.04	.75	.71	.04
		16	12	.99	.99	.00	1.00	1.00	.00	.98	.99	.01	.99	.99	.00	.99	.99	.00	.99	.99	.00	.99	1.00	.00
9	8	17	15	.86	.86	.00	.82	.85	.03	.85	.85	.00	.82	.86	.04	.73	.79	.06	.86	.87	.02	.86	.87	.01
		18	16	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	10	23	19	.84	.82	.02	.87	.87	.00	.82	.85	.03	.92	.90	.02	.84	.83	.01	.87	.86	.01	.85	.86	.01
		24	20	.91	.90	.00	.94	.93	.01	.91	.91	.00	.88	.87	.01	.89	.90	.00	.92	.92	.00	.91	.92	.01
13	13	25	25	.61	.59	.02	.64	.61	.02	.59	.60	.00	.61	.60	.01	.57	.54	.03	.60	.58	.01	.63	.60	.03
		26	26	.92	.91	.01	.90	.90	.00	.91	.89	.01	.92	.91	.01	.91	.90	.01	.90	.89	.01	.90	.91	.01
16	14	31	27	.80	.86	.05	.81	.87	.06	.82	.85	.02	.87	.92	.04	.83	.88	.04	.82	.86	.04	.83	.84	.02
		32	28	.80	.84	.04	.85	.89	.05	.86	.88	.01	.81	.85	.04	.79	.84	.06	.83	.87	.05	.82	.87	.05
17	18	33	35	.87	.84	.02	.87	.85	.03	.84	.80	.04	.86	.85	.01	.84	.81	.02	.85	.84	.01	.84	.83	.00
		34	36	.91	.90	.01	.94	.93	.01	.98	.95	.02	.95	.95	.00	.92	.91	.01	.94	.93	.01	.93	.94	.01
19	20	37	39	.60	.60	.00	.69	.65	.04	.67	.63	.03	.74	.75	.01	.61	.64	.02	.66	.65	.01	.64	.62	.02
		38	40	.59	.56	.02	.52	.49	.03	.60	.50	.10	.68	.64	.04	.58	.57	.02	.55	.51	.04	.55	.52	.02
22	21	43	41	.94	.94	.00	.97	.96	.01	.98	.94	.04	.95	.96	.01	.93	.93	.00	.96	.96	.00	.96	.96	.01
		44	42	.76	.76	.00	.84	.84	.00	.77	.74	.03	.78	.78	.00	.73	.74	.01	.82	.82	.00	.81	.81	.00
23	22	45	43	.89	.90	.01	.90	.91	.01	.88	.91	.03	.90	.90	.00	.87	.87	.00	.91	.92	.01	.91	.90	.01
		46	44	.93	.93	.00	.94	.93	.00	.91	.96	.05	.94	.94	.01	.92	.92	.00	.93	.95	.01	.94	.93	.01
25	23	49	45	.87	.86	.01	.87	.87	.00	.83	.83	.00	.87	.87	.00	.82	.81	.01	.88	.87	.01	.88	.86	.02
		50	46	.89	.89	.01	.91	.90	.01	.91	.88	.03	.92	.91	.00	.85	.86	.00	.91	.90	.01	.91	.90	.01

Note. NA indicates that the parameters could not be estimated, because these items did not have a correct LE response option.

APPENDIX K: Item-to-Competency Mapping

Item	T1		T2	
	Original Mapping	IST Mapping	Original Mapping	IST Mapping
1	Communication Skills	Leading	**Mentoring Others**	**Mentoring Others**
2	Decision-Making and Managing Resources	Decision Making	**Communication Skills**	**Communication**
3	**Mentoring Others**	**Mentoring Others**	**Mentoring Others**	**Mentoring Others**
4	**Communication Skills**	**Communication**	Communication Skills	Communication
5	**Mentoring Others**	**Mentoring Others**	Mentoring Others	Mentoring Others
6	Leading Others	Leading	**Leading Others**	**Leading**
7	Leading Innovation	Decision Making	Communication Skills	Communication
8	**Leading Others**	**Leading**	**Decision-Making and Managing Resources**	**Decision Making**
9	**Decision-Making and Managing Resources**	**Decision Making**	Mentoring Others	Mentoring Others
10	Displaying Integrity, Ethical Behavior, and Professionalism	Integrity	**Leading Others**	**Leading**
11	Leading Innovation	Innovation	Decision-Making and Managing Resources	Decision Making
12	**Leading Others**	**Leading**	Leading Innovation	Decision Making
13	**Communication Skills**	**Communication**	**Communication Skills**	**Communication**
14	Displaying Integrity, Ethical Behavior, and Professionalism	Integrity	**Leading Innovation**	**Innovation**
15	Displaying Integrity, Ethical Behavior, and Professionalism	Integrity	Displaying Integrity, Ethical Behavior, and Professionalism	Integrity
16	**Leading Innovation**	**Innovation**	Mentoring Others	Mentoring Others
17	**Leading Others**	**Leading**	Leading Others	Leading
18	Decision-Making and Managing Resources	Decision Making	**Leading Others**	**Leading**
19	**Communication Skills**	**Leading**	Communication Skills	Communication
20	Displaying Integrity, Ethical Behavior, and Professionalism	Integrity	**Communication Skills**	**Leading**
21	Mentoring Others	Mentoring Others	**Mentoring Others**	**Mentoring Others**
22	**Mentoring Others**	**Mentoring Others**	**Communication Skills**	**Integrity**
23	**Communication Skills**	**Integrity**	**Communication Skills**	**Communication**
24	Decision-Making and Managing Resources	Leading	Communication Skills	Communication
25	**Communication Skills**	**Communication**	Decision-Making and Managing Resources	Decision Making

Note. In red are the competencies for which there was a disagreement between the original and the IST mappings. In bold are the common items between T1 and T2. The disagreements were forwarded to the AFPC/DSYX for arbitration. The IST mapping was deemed final.

APPENDIX L. Stratified Samples Subtest-Level Descriptives

OTS-CIV

SUBTEST	T1					T2				
	N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
Overall	7,926	2.65	.56	-2.66	12.50	7,168	2.66	.54	-2.48	9.84
Most Effective	7,926	2.61	.62	-1.73	2.75	7,167	2.62	.59	-1.78	3.67
Least Effective	7,926	2.69	.50	-3.58	22.25	7,169	2.71	.50	-3.19	16.00
Integrity	7,910	2.70	.55	-3.71	27.94	7,104	2.54	.60	-1.66	3.30
Leading Others	7,931	2.66	.55	-2.71	13.24	7,209	2.72	.51	-3.60	25.19
Decision Making	8,001	2.51	.59	-2.04	5.49	7,127	2.58	.50	-2.32	7.97
Communication Skills	7,834	2.67	.58	-2.55	7.99	7,114	2.66	.59	-2.54	7.87
Leading Innovation	8,000	2.76	.57	-2.72	8.35	7,262	2.78	.53	-2.25	3.85
Mentoring Others	7,893	2.66	.54	-1.91	4.37	7,229	2.69	.52	-1.91	3.76

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

OTS-AD

SUBTEST	T1					T2				
	N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
Overall	5,388	2.70	.52	-3.22	19.68	4,793	2.70	.51	-2.84	13.13
Most Effective	5,388	2.67	.59	-2.05	4.72	4,792	2.65	.57	-2.04	5.71
Least Effective	5,389	2.73	.44	-4.39	34.64	4,794	2.75	.44	-3.63	20.55
Integrity	5,381	2.76	.51	-4.43	40.73	4,768	2.56	.59	-1.74	3.72
Leading Others	5,391	2.70	.50	-3.49	24.05	4,809	2.73	.50	-3.82	28.45
Decision Making	5,420	2.53	.57	-2.16	5.88	4,777	2.58	.48	-2.50	11.08
Communication Skills	5,352	2.70	.55	-2.98	12.82	4,773	2.71	.55	-3.03	12.51
Leading Innovation	5,419	2.79	.53	-3.09	11.66	4,829	2.84	.46	-2.67	5.92
Mentoring Others	5,373	2.75	.46	-2.53	8.67	4,815	2.76	.46	-2.41	6.78

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

AECP

SUBTEST	T1					T2				
	N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
Overall	169	2.68	.53	-2.62	9.75	165	2.67	.53	-2.46	8.45
Most Effective	169	2.64	.60	-1.96	5.05	165	2.62	.60	-1.68	2.83
Least Effective	169	2.72	.47	-3.31	14.66	165	2.73	.47	-3.23	14.06
Integrity	169	2.73	.53	-2.48	6.73	163	2.57	.55	-2.03	6.52
Leading Others	169	2.69	.52	-3.03	14.45	167	2.70	.53	-3.13	15.83
Decision Making	170	2.53	.57	-2.26	7.85	165	2.56	.51	-2.16	7.05
Communication Skills	168	2.66	.59	-2.45	7.31	164	2.67	.59	-2.58	7.93
Leading Innovation	170	2.77	.55	-2.60	6.90	168	2.81	.50	-2.40	4.60
Mentoring Others	169	2.73	.47	-2.55	10.11	167	2.73	.48	-2.11	5.11

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

USAFA

SUBTEST	T1					T2				
	N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
Overall	4,286	2.70	.52	-3.12	15.65	3,717	2.70	.50	-2.89	13.02
Most Effective	4,285	2.68	.57	-2.24	5.59	3,717	2.66	.56	-2.16	6.03
Least Effective	4,287	2.72	.46	-4.00	25.71	3,718	2.73	.45	-3.62	20.00
Integrity	4,286	2.77	.51	-3.75	22.33	3,709	2.50	.60	-1.74	4.26
Leading Others	4,288	2.69	.50	-3.45	21.80	3,726	2.80	.46	-4.03	25.59
Decision Making	4,296	2.56	.54	-2.75	11.83	3,705	2.59	.48	-2.68	13.28
Communication Skills	4,263	2.68	.57	-2.58	7.84	3,708	2.69	.56	-2.90	11.21
Leading Innovation	4,297	2.78	.55	-2.94	9.71	3,733	2.83	.48	-2.90	8.07
Mentoring Others	4,284	2.76	.48	-2.62	9.20	3,728	2.74	.47	-2.45	8.21

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

ROTC

SUBTEST	T1					T2				
	N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
Overall	12,623	2.63	.58	-2.50	10.32	12,244	2.63	.56	-2.31	8.12
Most Effective	12,623	2.59	.64	-1.62	2.33	12,241	2.58	.62	-1.65	3.24
Least Effective	12,624	2.67	.51	-3.37	18.31	12,246	2.69	.50	-2.98	13.00
Integrity	12,602	2.71	.56	-3.27	18.10	12,138	2.49	.63	-1.45	2.29
Leading Others	12,633	2.64	.55	-2.78	14.01	12,294	2.71	.53	-3.30	18.38
Decision Making	12,717	2.48	.61	-1.93	5.61	12,197	2.55	.51	-2.15	7.87
Communication Skills	12,504	2.60	.63	-2.11	5.18	12,177	2.62	.62	-2.27	6.37
Leading Innovation	12,716	2.74	.59	-2.48	6.15	12,370	2.80	.51	-2.43	4.89
Mentoring Others	12,583	2.64	.56	-1.91	4.76	12,325	2.66	.53	-1.91	4.28

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

ANG

SUBTEST	T1					T2				
	N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
Overall	3,890	2.69	.52	-3.04	15.79	3,260	2.70	.51	-2.73	11.08
Most Effective	3,889	2.66	.59	-2.00	4.24	3,260	2.65	.57	-2.04	5.72
Least Effective	3,890	2.72	.46	-4.09	27.33	3,261	2.74	.46	-3.42	16.44
Integrity	3,887	2.75	.52	-3.84	25.25	3,246	2.59	.56	-1.93	4.89
Leading Others	3,890	2.68	.52	-3.30	21.32	3,271	2.74	.50	-3.50	20.50
Decision Making	3,908	2.56	.56	-2.27	6.08	3,248	2.60	.48	-2.65	11.77
Communication Skills	3,865	2.70	.55	-3.04	14.09	3,247	2.69	.57	-2.86	10.68
Leading Innovation	3,907	2.78	.55	-3.00	11.08	3,279	2.81	.50	-2.40	4.46
Mentoring Others	3,884	2.73	.49	-2.40	7.60	3,276	2.75	.47	-2.32	6.40

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

AFRES

SUBTEST	T1					T2				
	N	*M*	*SD*	Skewness	Kurtosis	*N*	*M*	*SD*	Skewness	Kurtosis
Overall	1,468	2.69	.53	-2.92	13.94	1,202	2.69	.52	-2.79	14.30
Most Effective	1,468	2.66	.60	-1.98	4.26	1,202	2.64	.58	-1.94	5.19
Least Effective	1,468	2.72	.47	-3.87	23.63	1,202	2.74	.46	-3.63	23.41
Integrity	1,465	2.74	.53	-3.79	24.93	1,192	2.57	.59	-1.69	3.36
Leading Others	1,468	2.68	.53	-3.02	16.79	1,207	2.73	.50	-4.21	39.92
Decision Making	1,477	2.55	.56	-2.32	6.86	1,198	2.59	.48	-2.67	12.60
Communication Skills	1,460	2.70	.56	-2.75	9.57	1,195	2.68	.58	-2.69	8.92
Leading Innovation	1,478	2.78	.55	-2.84	8.94	1,215	2.80	.52	-2.29	3.93
Mentoring Others	1,466	2.73	.49	-2.44	8.11	1,208	2.74	.49	-2.25	5.74

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

APPENDIX M. Stratified Samples Subtest-Level Difficulty and Discriminability

OTS-CIV

SUBTEST	*p*-values						ITC					
	T1			T2			T1			T2		
	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*
Overall	.42	.99	.77	.32	.99	.76	-.08	.25	.14	-.03	.32	.17
Most Effective	.42	.94	.74	.32	.96	.73	-.08	.25	.13	-.03	.32	.16
Least Effective	.49	.99	.80	.42	.99	.79	.02	.23	.14	.07	.30	.17
Integrity	.52	.99	.81	.32	.93	.64	.09	.25	.14	.09	.22	.16
Leading Others	.42	.99	.78	.56	.99	.82	-.08	.23	.11	.05	.21	.13
Decision Making	.49	.96	.69	.52	.97	.70	.06	.23	.13	.08	.16	.14
Communication Skills	.47	.95	.79	.41	.96	.78	.09	.22	.16	-.02	.27	.15
Leading Innovation	.80	.97	.85	.84	.86	.85	.08	.13	.11	NA	NA	NA
Mentoring Others	.49	.94	.73	.37	.94	.75	.11	.21	.18	.08	.27	.16

Note. Min = Minimum; Max = Maximum; *M* = Mean. NA indicates that the parameters could not be estimated, because the Leading Innovation competency was measured by only 1 situation on T2.

OTS-AD

SUBTEST	*p*-values						ITC					
	T1			T2			T1			T2		
	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*
Overall	.44	1.00	.80	.37	1.00	.78	-.03	.29	.12	-.02	.37	.16
Most Effective	.52	.97	.78	.37	.98	.75	-.03	.26	.11	-.02	.37	.15
Least Effective	.44	1.00	.82	.39	1.00	.81	.02	.29	.13	.08	.34	.17
Integrity	.63	1.00	.86	.37	.93	.65	.04	.29	.13	.08	.24	.16
Leading Others	.44	1.00	.81	.49	1.00	.82	-.03	.18	.09	.05	.21	.13
Decision Making	.53	.96	.71	.51	.98	.70	.04	.19	.11	.07	.15	.11
Communication Skills	.52	.98	.80	.49	.98	.81	.10	.26	.16	-.03	.33	.14
Leading Innovation	.81	.98	.87	.87	.89	.88	.08	.12	.10	NA	NA	NA
Mentoring Others	.56	.97	.79	.55	.96	.80	.10	.26	.16	.07	.20	.13

Note. Min = Minimum; Max = Maximum; *M* = Mean. NA indicates that the parameters could not be estimated, because the Leading Innovation competency was measured by only 1 situation on T2.

AECP

SUBTEST	*p*-values						ITC					
	T1			T2			T1			T2		
	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*
Overall	.43	1.00	.77	.32	.99	.77	-.07	.31	.09	-.11	.37	.14
Most Effective	.48	.98	.73	.39	.95	.73	-.07	.27	.09	-.11	.37	.15
Least Effective	.43	1.00	.82	.32	.99	.80	-.07	.31	.09	-.06	.36	.14
Integrity	.56	1.00	.84	.32	.93	.64	-.02	.20	.07	.02	.22	.16
Leading Others	.43	.98	.73	.50	.99	.80	-.07	.26	.08	.05	.21	.13
Decision Making	.54	.97	.71	.51	.97	.69	-.07	.15	.05	-.01	.29	.09
Communication Skills	.48	.95	.78	.42	.96	.79	.09	.19	.14	-.09	.33	.11
Leading Innovation	.79	.96	.86	.85	.88	.86	.05	.12	.09	NA	NA	NA
Mentoring Others	.54	.98	.79	.50	.95	.78	.00	.31	.12	.05	.31	.16

Note. Min = Minimum; Max = Maximum; *M* = Mean. NA indicates that the parameters could not be estimated, because the Leading Innovation competency was measured by only 1 situation on T2.

USAFA

SUBTEST	*p*-values						ITC					
	T1			T2			T1			T2		
	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*
Overall	.36	.99	.81	.26	.99	.78	-.05	.22	.11	-.04	.26	.13
Most Effective	.41	.96	.80	.26	.97	.76	-.05	.19	.10	-.04	.26	.13
Least Effective	.36	.99	.82	.37	.99	.80	.00	.22	.12	.06	.21	.13
Integrity	.60	.99	.86	.26	.94	.62	.07	.22	.14	.13	.18	.15
Leading Others	.36	.99	.81	.64	.99	.86	-.05	.19	.10	.05	.21	.13
Decision Making	.51	.98	.73	.50	.99	.69	.05	.20	.10	.07	.19	.12
Communication Skills	.53	.95	.80	.51	.98	.80	.07	.18	.13	-.05	.21	.11
Leading Innovation	.81	.97	.87	.85	.92	.88	.07	.13	.09	NA	NA	NA
Mentoring Others	.60	.96	.80	.51	.96	.78	.06	.19	.12	.03	.21	.11

Note. Min = Minimum; Max = Maximum; *M* = Mean. NA indicates that the parameters could not be estimated, because the Leading Innovation competency was measured by only 1 situation on T2.

ROTC

SUBTEST	*p*-values						ITC					
	T1			T2			T1			T2		
	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*
Overall	.42	.99	.76	.32	.99	.74	-.04	.24	.14	-.06	.31	.15
Most Effective	.42	.93	.74	.35	.97	.71	-.04	.24	.13	-.06	.31	.15
Least Effective	.42	.99	.79	.32	.99	.78	-.02	.24	.14	.04	.26	.16
Integrity	.53	.99	.82	.32	.92	.62	.10	.22	.15	.13	.21	.17
Leading Others	.42	.99	.77	.57	.99	.81	-.04	.19	.11	.05	.21	.13
Decision Making	.46	.96	.68	.45	.97	.67	.04	.24	.13	.07	.16	.11
Communication Skills	.42	.94	.75	.39	.96	.76	.09	.24	.16	-.05	.27	.13
Leading Innovation	.79	.96	.84	.84	.88	.86	.08	.13	.11	NA	NA	NA
Mentoring Others	.45	.95	.73	.41	.95	.73	.10	.22	.17	.06	.27	.15

Note. Min = Minimum; Max = Maximum; *M* = Mean. NA indicates that the parameters could not be estimated, because the Leading Innovation competency was measured by only 1 situation on T2.

ANG

SUBTEST	*p*-values						ITC					
	T1			T2			T1			T2		
	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*
Overall	.46	.99	.80	.37	.99	.78	-.05	.26	.12	.04	.36	.17
Most Effective	.48	.96	.78	.37	.98	.75	-.05	.26	.11	.04	.36	.16
Least Effective	.46	.99	.82	.43	.99	.81	-.01	.25	.13	.06	.35	.18
Integrity	.60	.99	.85	.37	.95	.67	.05	.26	.13	.12	.24	.18
Leading Others	.46	.99	.80	.51	.99	.82	-.05	.20	.09	.05	.21	.13
Decision Making	.55	.96	.72	.54	.98	.71	.06	.18	.11	.05	.17	.12
Communication Skills	.54	.98	.80	.47	.98	.80	.06	.22	.13	.02	.28	.16
Leading Innovation	.82	.98	.86	.86	.87	.87	.07	.12	.10	NA	NA	NA
Mentoring Others	.56	.96	.78	.47	.96	.79	.06	.24	.16	.10	.26	.16

Note. Min = Minimum; Max = Maximum; *M* = Mean. NA indicates that the parameters could not be estimated, because the Leading Innovation competency was measured by only 1 situation on T2.

AFRES

SUBTEST	*p*-values						ITC					
	T1			T2			T1			T2		
	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*
Overall	.50	.99	.80	.37	1.00	.77	-.04	.30	.13	.02	.33	.16
Most Effective	.50	.96	.77	.37	.98	.74	-.04	.25	.13	.02	.33	.15
Least Effective	.50	.99	.82	.44	1.00	.81	-.01	.30	.13	.06	.32	.17
Integrity	.57	.99	.84	.37	.93	.66	.03	.30	.15	.05	.24	.13
Leading Others	.50	.99	.79	.52	1.00	.82	-.04	.25	.10	.05	.21	.13
Decision Making	.52	.96	.72	.53	.98	.70	.06	.23	.12	.06	.20	.11
Communication Skills	.53	.96	.80	.45	.97	.80	.08	.28	.16	.02	.31	.15
Leading Innovation	.82	.97	.86	.84	.87	.86	.08	.13	.10	NA	NA	NA
Mentoring Others	.59	.96	.78	.49	.96	.78	.08	.29	.17	.09	.26	.15

Note. Min = Minimum; Max = Maximum; *M* = Mean. NA indicates that the parameters could not be estimated, because the Leading Innovation competency was measured by only 1 situation on T2.

APPENDIX N. Stratified Samples Subtest-Level Internal Consistency

SUBTEST	OTS-CIV				OTS-AD				AECP				USAFA			
	T1		T2		T1		T2		T1		T2		T1		T2	
	α	*ω*	*α*	*ω*	*α*	*ω*	*α*	*ω*	*α*	*ω*	*α*	*ω*	*α*	*ω*	*α*	*ω*
Overall	.56	.67	.65	.72	.52	.62	.63	.71	.45	NA	.59	NA	.48	.61	.54	.65
Most Effective	.36	.48	.46	.54	.32	.41	.43	.51	.32	NA	.38	NA	.27	.39	.34	.46
Least Effective	.40	.53	.50	.57	.38	.49	.50	.51	.32	NA	.40	NA	.34	NA	.39	NA
Integrity	.29	NA	.29	NA	.31	NA	.29	NA	.17	NA	.27	NA	.31	NA	.29	NA
Leading Others	.19	NA	.30	NA	.15	NA	.30	NA	.16	NA	.28	NA	.18	NA	.29	NA
Decision Making	.35	NA	.31	NA	.30	NA	.26	NA	.15	NA	.22	NA	.30	NA	.27	NA
Communication Skills	.27	NA	.38	NA	.31	NA	.37	NA	.37	NA	.28	NA	.25	NA	.29	NA
Leading Innovation	.22	NA	NA	NA	.18	NA	NA	NA	.06	NA	NA	NA	.18	NA	NA	NA
Mentoring Others	.37	NA	.42	NA	.33	NA	.35	NA	.20	NA	.41	NA	.24	NA	.30	NA

SUBTEST	ROTC				ANG				AFRES			
	T1		T2		T1		T2		T1		T2	
	α	*ω*	*α*	*ω*	*α*	*ω*	*α*	*ω*	*α*	*ω*	*α*	*ω*
Overall	.56	.67	.65	.72	.52	.62	.63	.71	.45	NA	.59	NA
Most Effective	.36	.48	.46	.54	.32	.41	.43	.51	.32	NA	.38	NA
Least Effective	.40	.53	.50	.57	.38	.49	.50	.51	.32	NA	.40	NA
Integrity	.29	NA	.29	NA	.31	NA	.29	NA	.17	NA	.27	NA
Leading Others	.19	NA	.30	NA	.15	NA	.30	NA	.16	NA	.28	NA
Decision Making	.35	NA	.31	NA	.30	NA	.26	NA	.15	NA	.22	NA
Communication Skills	.27	NA	.38	NA	.31	NA	.37	NA	.37	NA	.28	NA
Leading Innovation	.22	NA	NA	NA	.18	NA	NA	NA	.06	NA	NA	NA
Mentoring Others	.37	NA	.42	NA	.33	NA	.35	NA	.20	NA	.41	NA

Note. *α* = Cronbach's alpha; *ω* = McDonald's omega. NA indicates that the parameters could not be estimated due to factorial complexity of the subtests and due to the fact that the Leading Innovation competency was measured by only 1 situation on T2.

APPENDIX O. Stratified Samples Subtest-Level Subgroup Differences

OTS-CIV

SUBTEST	T1 Cohen's *d**					T2 Cohen's *d**				
	F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
Overall	.10	.72	.51	.22	.16	.02	.72	.61	.35	.05
Most Effective	.11	.65	.36	.24	.13	.06	.67	.35	.32	.07
Least Effective	.06	.54	.42	.12	.15	.10	.55	.45	.28	.02
Integrity	.10	.44	.23	.14	.04	.11	.27	.27	.17	.04
Leading Others	.04	.29	.12	.06	.14	.04	.58	.33	.19	.16
Decision Making	.18	.45	.36	.19	.32	.06	.41	.27	.15	.05
Communication Skills	.01	.39	.13	.10	.08	.03	.42	.14	.25	.08
Leading Innovation	.11	.26	.26	.04	.01	.10	.14	.14	.18	.02
Mentoring Others	.03	.41	.36	.14	.02	.02	.41	.37	.22	.13

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.
Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic.

OTS-AD

SUBTEST	T1 Cohen's *d**					T2 Cohen's *d**				
	F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
Overall	.12	.65	.47	.20	.01	.03	.75	.55	.34	.02
Most Effective	.15	.54	.36	.19	.01	.11	.68	.35	.27	.02
Least Effective	.05	.53	.42	.13	.00	.08	.60	.45	.32	.06
Integrity	.11	.30	.23	.14	.10	.08	.38	.27	.15	.08
Leading Others	.05	.23	.12	.04	.13	.05	.62	.33	.30	.04
Decision Making	.16	.33	.36	.06	.11	.01	.28	.27	.08	.03
Communication Skills	.07	.40	.13	.11	.02	.03	.47	.14	.19	.03
Leading Innovation	.09	.24	.26	.14	.11	.06	.11	.14	.14	.06
Mentoring Others	.02	.51	.36	.16	.15	.05	.47	.37	.21	.15

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.
Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic.

AECP

SUBTEST	T1 Cohen’s *d**					T2 Cohen’s *d**				
	F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
Overall	.13	.70	.02	.23	.30	.46	.88	.89	.38	.88
Most Effective	.21	.55	.36	.12	.49	.50	.88	.35	.29	.51
Least Effective	.02	.59	.42	.28	.02	.27	.56	.45	.34	1.15
Integrity	.01	.42	.23	.08	.06	.06	.43	.27	.04	.36
Leading Others	.25	.37	.12	.43	.52	.44	.60	.33	.17	.79
Decision Making	.06	.02	.36	.08	.44	.25	.65	.27	.34	.17
Communication Skills	.17	.53	.13	.11	.27	.19	.74	.14	.60	.27
Leading Innovation	.08	.71	.26	.03	.03	.07	.41	.14	.35	.41
Mentoring Others	.19	.31	.36	.02	.63	.30	.31	.37	.29	.63

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic.

USAFA

SUBTEST	T1 Cohen’s *d**					T2 Cohen’s *d**				
	F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
Overall	.00	.56	.25	.06	.19	.12	.36	.22	.08	.06
Most Effective	.01	.53	.36	.04	.27	.07	.25	.35	.07	.15
Least Effective	.00	.39	.42	.06	.06	.13	.38	.45	.07	.27
Integrity	.03	.30	.23	.01	.17	.07	.10	.27	.18	.11
Leading Others	.06	.07	.12	.01	.10	.02	.33	.33	.02	.11
Decision Making	.18	.34	.36	.06	.02	.11	.25	.27	.01	.18
Communication Skills	.03	.37	.13	.01	.19	.02	.13	.14	.06	.05
Leading Innovation	.15	.24	.26	.02	.05	.13	.16	.14	.07	.22
Mentoring Others	.02	.43	.36	.14	.12	.11	.20	.37	.00	.01

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic.

ROTC

SUBTEST	T1 Cohen’s *d**					T2 Cohen’s *d**				
	F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
Overall	.13	.75	.41	.28	.12	.02	.79	.38	.36	.10
Most Effective	.12	.72	.36	.26	.12	.07	.72	.35	.30	.07
Least Effective	.10	.52	.42	.21	.08	.05	.63	.45	.32	.11
Integrity	.03	.43	.23	.22	.03	.09	.26	.27	.17	.01
Leading Others	.10	.28	.12	.03	.08	.12	.71	.33	.26	.11
Decision Making	.21	.36	.36	.20	.05	.03	.26	.27	.14	.06
Communication Skills	.05	.50	.13	.17	.03	.03	.50	.14	.19	.03
Leading Innovation	.08	.35	.26	.09	.07	.11	.15	.14	.10	.03
Mentoring Others	.04	.48	.36	.17	.20	.00	.52	.37	.27	.16

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.
Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic.

ANG

SUBTEST	T1 Cohen’s *d**					T2 Cohen’s *d**				
	F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
Overall	.03	.63	.56	.21	.16	.00	.63	.26	.22	.07
Most Effective	.05	.63	.36	.17	.12	.06	.58	.35	.18	.08
Least Effective	.00	.41	.42	.18	.15	.08	.51	.45	.21	.04
Integrity	.02	.29	.23	.09	.29	.08	.25	.27	.04	.04
Leading Others	.08	.30	.12	.02	.00	.02	.54	.33	.18	.19
Decision Making	.12	.34	.36	.12	.17	.01	.31	.27	.07	.06
Communication Skills	.08	.32	.13	.13	.32	.01	.44	.14	.15	.12
Leading Innovation	.12	.27	.26	.12	.00	.09	.13	.14	.05	.22
Mentoring Others	.00	.46	.36	.19	.23	.03	.36	.37	.24	.14

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.
Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic.

AFRES

SUBTEST	T1 Cohen's *d**					T2 Cohen's *d**				
	F/M	B/W	A/W	WnH/WH	BnH/BH	F/M	B/W	A/W	WnH/WH	BnH/BH
Overall	.12	.85	.48	.26	.01	.03	.90	.45	.21	.11
Most Effective	.16	.71	.36	.14	.10	.13	.82	.35	.33	.07
Least Effective	.04	.72	.42	.30	.14	.10	.69	.45	.01	.28
Integrity	.18	.48	.23	.20	.04	.03	.51	.27	.11	.07
Leading Others	.04	.38	.12	.01	.14	.05	.70	.33	.19	.10
Decision Making	.22	.44	.36	.10	.27	.05	.41	.27	.17	.16
Communication Skills	.03	.40	.13	.11	.21	.05	.62	.14	.06	.09
Leading Innovation	.19	.28	.26	.15	.05	.11	.04	.14	.10	.19
Mentoring Others	.01	.61	.36	.27	.21	.03	.56	.37	.06	.26

*Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

Note. F/M = Female/Male; B/W = Black/White; A/W = Asian/White; WnH/WH = White Non-Hispanic/White Hispanic; BnH/BH = Black Non-Hispanic/Black Hispanic.

APPENDIX P. Unstratified Sample Subtest-Level SJT Correlations with SDI-O

NOTE: A data quality issue with the SDI-O was discovered after data analysis was completed and after the results were presented to the AFPC/DSYX. This data quality issue involved the High Intensity Pleasure, Deliberation, Spontaneous Variety, and Imagination facets and in turn the SDI-O domains, OSM facets, and OSM score. Reported below are a description of the scope of the data quality problem, the original correlations using the inaccurate data, and a new set of correlations. The new set of correlations includes the correlations with the problematic facets and affected SDI-O domains, OSM facets, and OSM score re-run with accurate data. This did drastically affect sample size (*N* = 22,794 [T1] and *N* = 20,940 [T2] in the new dataset compared to *N* = 36,099 [T1] and *N* = 32,952 [T2] in the original dataset). Results here are presented for historical documentation only and should be interpreted with caution.

Problematic Facets

FACETS	Number of Problematic Data Points	Percentage of Test-Takers with at Least One Problematic Data Point	Correlation Between Corrected Data and Uncorrected Data
High-Intensity Pleasure	8877	14.04%	.99
Deliberation	11064	17.5%	.97
Spontaneous Variety	7624	12.06%	.99
Imagination	8582	13.58%	.99

T1 Original Correlations

Name	Overall	ME	LE	Integ.	Lead.	Dec.	Comm.	Inn.	Ment.
FACETS									
Team Player	.07	.05	.07	.04	.03	.04	.03	.05	.05
Stress Under Pressure	-.13	-.11	-.12	-.06	-.07	-.09	-.07	-.04	-.09
Reserved	-.06	-.04	-.07	-.03	-.03	-.04	-.02	-.03	-.04
Achievement Striving	.05	.03	.06	.02	.03	.03	.02	.03	.03
Creative	.01	-.01	.03	-.02	.00	.03	.00	.00	.01
Interpersonal Tactics	-.11	-.09	-.10	-.05	-.07	-.04	-.05	-.08	-.08
Pleasant	.01	-.01	.02	.01	.01	-.01	.00	.01	.00
Temperamental	-.06	-.04	-.06	-.02	-.03	-.04	-.03	-.03	-.04
Dominance - Leader	.06	.04	.07	.03	.02	.05	.02	.02	.05
Order	-.03	-.04	.00	-.01	.00	-.04	-.02	.00	-.02
Reflective	-.05	-.05	-.03	-.05	-.01	-.01	-.04	-.01	-.03
Cynical View	-.10	-.08	-.10	-.04	-.05	-.06	-.06	-.05	-.08
Helpful Altruistic	.01	.00	.02	.00	.01	-.01	.00	.02	.02
Worry	-.11	-.09	-.09	-.04	-.05	-.08	-.07	-.05	-.08
Excitement Seeking	.02	.02	.02	.01	.00	.03	.01	.01	.01
Self-Discipline	.04	.01	.05	.01	.02	.00	.03	.02	.02
Scientific Interest	.04	.03	.03	.00	.01	.04	.04	.01	.02
Envious	.01	.02	-.01	.02	-.02	.02	.01	-.01	.01
Independent	-.01	-.01	-.02	-.01	-.01	.00	-.01	-.01	-.01
Angry Hostile	-.05	-.04	-.05	-.02	-.02	-.04	-.03	-.03	-.03
High Intensity	-.01	-.01	.00	.00	.00	.00	-.02	.00	.00
Deliberation	.00	-.01	.02	-.01	.02	.00	-.01	.00	.00
Cultured	-.04	-.05	-.02	-.06	-.01	-.03	-.03	.00	-.02
Influence Tactics	.04	.04	.03	.01	.00	.07	.03	.01	.03
Optimist	.02	.00	.03	.00	.01	.00	.01	.01	.02
Unconventional	-.01	.01	-.03	-.01	-.03	.01	.00	-.02	.00
Spontaneous-Variety	-.07	-.06	-.05	-.04	-.04	-.04	-.04	-.03	-.03
Activity	.03	.01	.04	.02	.00	.02	.03	.01	.04
Well Adjusted	.08	.06	.08	.05	.03	.04	.04	.05	.06
Imagination	-.01	-.01	.00	-.01	-.01	.00	-.01	.01	.00
DOMAINS									
Agreeableness	.04	.02	.05	.02	.02	.00	.01	.03	.03
Conscientiousness	.02	.00	.04	.00	.02	.00	.01	.02	.01
Extraversion	.03	.01	.04	.01	.00	.03	.01	.02	.03
Neuroticism	-.11	-.09	-.10	-.04	-.05	-.08	-.06	-.05	-.07
Openness	-.01	-.02	.00	-.04	.00	.01	-.01	.00	.00
Machiavellianism	-.05	-.03	-.06	-.02	-.04	-.01	-.03	-.04	-.04
OSM FACETS									
Team Player	.07	.05	.07	.04	.03	.04	.03	.05	.05
Stress Under Pressure	-.13	-.11	-.12	-.06	-.07	-.09	-.07	-.04	-.09
Unassertive	-.06	-.04	-.07	-.03	-.03	-.04	-.02	-.03	-.04
Hyper-Competitive	-.10	-.08	-.09	-.05	-.06	-.03	-.05	-.07	-.07
Dominance-Leader	.04	.03	.05	.02	.01	.05	.02	.01	.04
OSM	.08	.06	.08	.04	.05	.04	.04	.04	.06

Note. ME = Most Effective; LE = Least Effective; Integ. = Integrity; Lead. = Leading Others; Dec. = Decision Making; Comm. = Communication Skills; Inn. = Leading Innovation; Ment. = Mentoring Others. *N* = 36,078.

T1 Corrected Correlations

Name	Overall	ME	LE	Integ.	Lead.	Dec.	Comm.	Inn.	Ment.
FACETS									
High Intensity	.00	-.01	.00	.00	.00	.00	-.02	.01	.00
Deliberation	.00	-.02	.02	-.01	.01	.01	-.01	.00	-.01
Spontaneous-Variety	-.07	-.06	-.05	-.04	-.04	-.03	-.04	-.02	-.03
Imagination	-.01	-.01	.00	-.01	-.01	.00	-.01	.02	.00
DOMAINS									
Conscientiousness	.01	-.01	.03	.00	.02	.00	.00	.01	.00
Extraversion	.02	.00	.03	.01	-.01	.02	.00	.01	.03
Openness	-.02	-.03	.00	-.04	-.01	.02	-.02	.00	.00

Note. ME = Most Effective; LE = Least Effective; Integ. = Integrity; Lead. = Leading Others; Dec. = Decision Making; Comm. = Communication Skills; Inn. = Leading Innovation; Ment. = Mentoring Others. *N* = 22,794.

T2 Original Correlations

Name	Overall	ME	LE	Integ.	Lead.	Dec.	Comm.	Inn.	Ment.
FACETS									
Team Player	.08	.06	.09	.05	.04	.06	.04	.02	.08
Stress Under Pressure	-.13	-.11	-.11	-.08	-.06	-.08	-.09	.00	-.11
Reserved	-.06	-.04	-.07	-.05	.00	-.06	-.03	-.01	-.06
Achievement Striving	.06	.04	.07	.05	.01	.06	.04	.01	.06
Creative	.00	.00	.01	.01	-.02	.01	-.01	.00	.02
Interpersonal Tactics	-.13	-.11	-.12	-.07	-.07	-.07	-.07	-.03	-.11
Pleasant	.01	-.01	.02	.01	-.04	.03	.00	.01	.02
Temperamental	-.07	-.05	-.07	-.05	-.02	-.05	-.04	-.01	-.06
Dominance - Leader	.05	.03	.05	.06	.00	.05	.03	.01	.04
Order	-.03	-.05	.00	.00	-.07	.02	-.02	.00	-.01
Reflective	-.05	-.04	-.05	-.05	-.03	-.03	-.05	.00	-.02
Cynical View	-.11	-.10	-.10	-.06	-.05	-.06	-.08	-.01	-.10
Helpful Altruistic	.01	.00	.03	.01	-.02	.02	-.01	.02	.04
Worry	-.11	-.10	-.09	-.06	-.06	-.05	-.08	-.01	-.10
Excitement Seeking	.03	.03	.03	.00	.04	.03	.01	.02	.01
Self-Discipline	.05	.03	.06	.05	-.02	.05	.04	.00	.05
Scientific Interest	.02	.03	.00	.00	.03	-.01	.03	-.01	.02
Envious	.00	.01	-.02	-.03	.04	-.02	.00	.01	-.02
Independent	-.01	-.01	-.01	-.02	.01	-.02	.00	.00	-.02
Angry Hostile	-.06	-.05	-.06	-.05	-.02	-.04	-.04	-.01	-.06
High Intensity	-.02	-.03	.00	-.02	-.02	.00	-.03	.01	.00
Deliberation	.00	-.01	.02	.01	-.03	.02	.00	.00	.02
Cultured	-.04	-.04	-.03	-.02	-.05	-.02	-.04	.00	.00
Influence Tactics	.03	.03	.02	.02	.02	.03	.02	.01	.02
Optimist	.01	-.01	.03	.02	-.03	.02	.00	.01	.03
Unconventional	-.02	.00	-.03	-.04	.03	-.04	.00	.01	-.02
Spontaneous-Variety	-.05	-.05	-.04	-.03	-.04	-.01	-.04	.01	-.03
Activity	.05	.03	.06	.05	.00	.05	.02	.01	.04
Well Adjusted	.09	.07	.09	.05	.05	.06	.04	.04	.08
Imagination	.00	.01	-.01	-.03	.03	-.02	-.01	.02	.01
DOMAINS									
Agreeableness	.04	.02	.05	.03	-.01	.04	.01	.02	.06
Conscientiousness	.03	.00	.05	.04	-.03	.05	.01	.00	.04
Extraversion	.03	.02	.05	.03	.00	.05	.01	.02	.03
Neuroticism	-.11	-.10	-.10	-.07	-.05	-.07	-.08	-.01	-.10
Openness	-.02	-.01	-.02	-.03	-.01	-.03	-.02	.01	.01
Machiavellianism	-.07	-.05	-.07	-.05	-.02	-.04	-.04	-.01	-.07
OSM FACETS									
Team Player	.08	.06	.09	.05	.04	.06	.04	.02	.08
Stress Under Pressure	-.13	-.11	-.11	-.08	-.06	-.08	-.09	.00	-.11
Unassertive	-.06	-.04	-.07	-.05	.00	-.06	-.03	-.01	-.06
Hyper-Competitive	-.11	-.09	-.10	-.06	-.06	-.06	-.06	-.02	-.10
Dominance-Leader	.04	.03	.04	.04	-.01	.04	.02	.00	.04
OSM	.08	.07	.07	.05	.03	.04	.05	.02	.08

Note. ME = Most Effective; LE = Least Effective; Integ. = Integrity; Lead. = Leading Others; Dec. = Decision Making; Comm. = Communication Skills; Inn. = Leading Innovation; Ment. = Mentoring Others. *N* = 32,932.

T2 Corrected Correlations

Name	Overall	ME	LE	Integ.	Lead.	Dec.	Comm.	Inn.	Ment.
FACETS									
High Intensity	.00	-.01	.00	.00	.00	.00	-.02	.01	.00
Deliberation	.00	-.02	.02	-.01	.01	.01	-.01	.00	-.01
Spontaneous-Variety	-.07	-.06	-.05	-.04	-.04	-.03	-.04	-.02	-.03
Imagination	-.01	-.01	.00	-.01	-.01	.00	-.01	.02	.00
DOMAINS									
Conscientiousness	.01	-.01	.03	.00	.02	.00	.00	.01	.00
Extraversion	.02	.00	.03	.01	-.01	.02	.00	.01	.03
Openness	-.02	-.03	.00	-.04	-.01	.02	-.02	.00	.00

Note. ME = Most Effective; LE = Least Effective; Integ. = Integrity; Lead. = Leading Others; Dec. = Decision Making; Comm. = Communication Skills; Inn. = Leading Innovation; Ment. = Mentoring Others. *N* = 20,940.

APPENDIX Q. Stratified Samples Subtest-Level Test-Retest Reliability

	OTS-CIV		OTS-AD		AECP		USAFA		ROTC		ANG		AFRES	
SUBTEST	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
	r_{T1T1}	r_{T2T2}	r_{T1T1}	r_{T2T2}	r_{T1T1}	r_{T2T2}	r_{T1T1}	r_{T2T2}	r_{T1T1}	r_{T2T2}	r_{T1T1}	r_{T2T2}	r_{T1T1}	r_{T2T2}
Overall	.77	.74	.75	.81	NA	NA	.77	NA	.52	.57	.73	.75	.56*	.45*
Most Effective	.80	.76	.76	.82	NA	NA	.85	NA	.47	.51	.67	.71	.60*	.45*
Least Effective	.66	.75	.61	.75	NA	NA	.78	NA	.52	.52	.57	.64	.41†	.42†
Integrity	.75	.82	.55	.65	NA	NA	.91	NA	.48	.43	.50	.74	.66	.52*
Leading Others	.69	.78	.52	.71	NA	NA	.66	NA	.34	.53	.69	.70	.58*	.06†
Decision Making	.68	.70	.68	.52	NA	NA	.83	NA	.32	.27	.66	.32*	.70	.24†
Communication Skills	.67	.80	.48	.66	NA	NA	.86	NA	.44	.41	.67	.75	.39†	.36†
Leading Innovation	.51	.80	.66	.65	NA	NA	.89	NA	.43	.32	.44	.32	-.06†	.43†
Mentoring Others	.60	.66	.66	.73	NA	NA	.84	NA	.53	.48	.73	.72	.73	.68

$^{*}p \leq .05$; $^{\dagger}p$ not significant; the rest of the correlations are significant at $p \leq .01$.

Note. r = Correlation. NA indicates that the parameters could not be estimated due to a small samples size. OTS-CIV: $N_{\mathrm{T1}} = 194$, $N_{\mathrm{T2}} = 146$; OTS-AD: $N_{\mathrm{T1}} = 200$, $N_{\mathrm{T2}} = 166$; AECP: $N_{\mathrm{T1}} = 0$, $N_{\mathrm{T2}} = 2$; USAFA: $N_{\mathrm{T1}} = 162$, $N_{\mathrm{T2}} = 16$; ROTC: $N_{\mathrm{T1}} = 702$, $N_{\mathrm{T2}} = 676$; ANG: $N_{\mathrm{T1}} = 126$, $N_{\mathrm{T2}} = 82$; AFRES: $N_{\mathrm{T1}} = 38$, $N_{\mathrm{T2}} = 40$.

APPENDIX R. Stratified Samples Subtest-Level Alternate Forms Reliability

OTS-CIV

SUBTEST	All Items Between Subjects	Common Items Between Subjects	All Items Within Subjects	
	Effect Size[1]	Effect Size[1]	Effect Size[1]	r_{T1T2}[2]
Overall	.07	.06	.09	.54
Most Effective	.01	.09	.03	.48
Least Effective	.12	.02	.13	.40
Integrity	.55	.01	.35	.06†
Leading Others	.37	.07	.14	.32
Decision Making	.29	.07	.32	.28
Communication Skills	.09	.10	.01	.25
Leading Innovation	.06	.15	.13	.17
Mentoring Others	.12	.04	.09	.31

[1] Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

[2] †*p* not significant; the rest of the correlations are significant at $p \leq .01$.

Note. *r* = Correlation; Effect Size = Cohen's *d*. All Items Between Subjects *N* = 14,597*;* Common Items Between Subjects *N* = 14,597; All Items Within Subjects *N* = 738.

OTS-AD

SUBTEST	All Items Between Subjects	Common Items Between Subjects	All Items Within Subjects	
	Effect Size[1]	Effect Size[1]	Effect Size[1]	r_{T1T2}[2]
Overall	.03	.05	.04	.48
Most Effective	.13	.08	.11	.38
Least Effective	.10	.00	.05	.41
Integrity	.73	.01	.57	.23
Leading Others	.24	.12	.07	.22
Decision Making	.19	.00	.10	.10*
Communication Skills	.05	.10	.12	.22
Leading Innovation	.14	.19	.13	.18
Mentoring Others	.05	.02	.04	.28

[1] Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

[2] $p \leq .05$; the rest of the correlations are significant at $p \leq .01$.

Note. *r* = Correlation; Effect Size = Cohen's *d*. All Items Between Subjects *N* = 9,804*;* Common Items Between Subjects *N* = 9,804; All Items Within Subjects *N* = 1,290.

AECP

SUBTEST	All Items Between Subjects	Common Items Between Subjects	All Items Within Subjects	
	Effect Size[1]	Effect Size[1]	Effect Size[1]	r_{T1T2}[2]
Overall	$.11^{\dagger}$	$.10^{\dagger}$	NA	NA
Most Effective	$.20^{\dagger}$	$.16^{\dagger}$	NA	NA
Least Effective	$.04^{\dagger}$	$.00^{\dagger}$	NA	NA
Integrity	.64**	$.20^{\dagger}$	NA	NA
Leading Others	$.12^{\dagger}$	.22*	NA	NA
Decision Making	$.13^{\dagger}$	$.03^{\dagger}$	NA	NA
Communication Skills	$.02^{\dagger}$	$.12^{\dagger}$	NA	NA
Leading Innovation	$.13^{\dagger}$	$.07^{\dagger}$	NA	NA
Mentoring Others	$.05^{\dagger}$	$.06^{\dagger}$	NA	NA

[1] Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

[2] $^{\dagger}p$ not significant; $*p \le .05$; the rest of the correlations are significant at $p \le .01$.

Note. r = Correlation; Effect Size = Cohen's d. NA indicates that the parameters could not be estimated due to a small samples size. All Items Between Subjects N = 323; Common Items Between Subjects N = 323; All Items Within Subjects N = 12.

USAFA

SUBTEST	All Items Between Subjects	Common Items Between Subjects	All Items Within Subjects	
	Effect Size[1]	Effect Size[1]	Effect Size[1]	r_{T1T2}[2]
Overall	.04	.01	.12	.33
Most Effective	.15	.03	.22	.26
Least Effective	.10	.00	.05	.24
Integrity	.96	.01	.72	$.08^{\dagger}$
Leading Others	.63	.05	.37	.19
Decision Making	.14	.02	.08	.11*
Communication Skills	.06	.07	.15	.21
Leading Innovation	.15	.16	.14	.20
Mentoring Others	.09	.01	.12	.20

[1] Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

[2] $^{\dagger}p$ not significant; $*p \le .05$; the rest of the correlations are significant at $p \le .01$.

Note. r = Correlation; Effect Size = Cohen's d. All Items Between Subjects N = 7,653; Common Items Between Subjects N = 7,653; All Items Within Subjects N = 804.

ROTC

SUBTEST	All Items Between Subjects	Common Items Between Subjects	All Items Within Subjects	
	Effect Size[1]	Effect Size[1]	Effect Size[1]	r_{T1T2}[2]
Overall	.01	.02	.01	.49
Most Effective	.09	.07	.09	.39
Least Effective	.12	.03	.09	.38
Integrity	.72	.01	.56	.18
Leading Others	.34	.06	.14	.27
Decision Making	.28	.06	.27	.13
Communication Skills	.00	.10	.01	.24
Leading Innovation	.15	.17	.12	.14
Mentoring Others	.10	.00	.10	.35

[1] Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

[2] All correlations are significant at $p \leq .01$.

Note. r = Correlation; Effect Size = Cohen's d. All Items Between Subjects N = 24,028*;* Common Items Between Subjects N = 24,028; All Items Within Subjects N = 5,842.

ANG

SUBTEST	All Items Between Subjects	Common Items Between Subjects	All Items Within Subjects	
	Effect Size[1]	Effect Size[1]	Effect Size[1]	r_{T1T2}[2]
Overall	.01	.04	.37	.00
Most Effective	.10	.08	.29	.00
Least Effective	.13	.00	.36	.00
Integrity	.60	.04	.09	.07
Leading Others	.36	.07	.27	.00*
Decision Making	.15	.02	.10	.04†
Communication Skills	.09	.10	.18	.00†
Leading Innovation	.10	.14	.15	.00†
Mentoring Others	.06	.00	.28	.00†

[1] Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

[2] †p not significant; *$p \leq .05$; the rest of the correlations are significant at $p \leq .01$.

Note. r = Correlation; Effect Size = Cohen's d. All Items Between Subjects N = 6,875*;* Common Items Between Subjects N = 6,875; All Items Within Subjects N = 884.

AFRES

SUBTEST	All Items Between Subjects	Common Items Between Subject	All Items Within Subjects	
	Effect Size[1]	Effect Size[1]	Effect Size[1]	r_{T1T2}[2]
Overall	.03	.05	.10	.36
Most Effective	.12	.08	.18	.37
Least Effective	.09	.00	.02	.27
Integrity	.63	.04	.44	.11†
Leading Others	.33	.05	.15	.36
Decision Making	.18	.03	.10	.09†
Communication Skills	.14	.09	.08	.29
Leading Innovation	.06	.13	.21	.03†
Mentoring Others	.01	.05	.10	.21

[1] Significance levels are available in Attachment 4 – SJT Item- and Subtest-Level Analyses.

[2] †*p* not significant; the rest of the correlations are significant at $p \leq .01$.

Note. *r* = Correlation; Effect Size = Cohen's *d*. All Items Between Subjects *N* = 2,584; Common Items Between Subjects *N* = 2,584; All Items Within Subjects *N* = 362.

APPENDIX S. Stratified Samples Retesting Effects on the Same and Parallel Forms

SUBTEST	OTS-CIV (Effect Size)			OTS-AD (Effect Size)			AECP (Effect Size)			USAFA (Effect Size)		
	T1	T2	T1/T2 or T2/T1	T1	T2	T1/T2 or T2/T1	T1	T2	T1/T2 or T2/T1	T1	T2	T1/T2 or T2/T1
Overall	.25*	.10†	.15**	.23*	.19†	.14**	NA	NA	NA	.15†	NA	.28**
Most Effective	.15†	.01†	.09†	.20*	.17†	.10**	NA	NA	NA	.13†	NA	.20**
Least Effective	.25*	.17†	.15**	.19†	.16†	.14**	NA	NA	NA	.15†	NA	.25**
Integrity	.01†	.00†	.11*	.09†	.13†	.10*	NA	NA	NA	.01†	NA	.17**
Leading Others	.16†	.08†	.05†	.09†	.01†	.04†	NA	NA	NA	.22*	NA	.05†
Decision Making	.20*	.13†	.01†	.06†	.10†	.01†	NA	NA	NA	.17†	NA	.09†
Communication Skills	.09†	.04†	.20**	.31**	.05†	.11**	NA	NA	NA	.01†	NA	.14**
Leading Innovation	.03†	.20†	.03†	.14†	.03†	.02†	NA	NA	NA	.11†	NA	.11*
Mentoring Others	.16†	.17†	.16**	.27**	.33**	.12**	NA	NA	NA	.08†	NA	.19**

SUBTEST	ROTC (Effect Size)			ANG (Effect Size)			AFRES (Effect Size)		
	T1	T2	T1/T2 or T2/T1	T1	T2	T1/T2 or T2/T1	T1	T2	T1/T2 or T2/T1
Overall	.22*	.13*	.18**	.04†	.04†	.02†	.09†	.19†	.06†
Most Effective	.13*	.08†	.14**	.03†	.14†	.00†	.14†	.16†	.03†
Least Effective	.25**	.14**	.13**	.03†	.09†	.05†	.01†	.17†	.13†
Integrity	.19**	.04†	.08**	.08†	.09†	.07†	.05†	.34†	.02†
Leading Others	.01†	.09†	.03†	.13†	.13†	.01†	.26†	.10†	.06†
Decision Making	.12*	.03†	.07**	.02†	.18†	.01†	.05†	.23†	.07†
Communication Skills	.16**	.13*	.10**	.18†	.14†	.07†	.04†	.17†	.00†
Leading Innovation	.02†	.05†	.07**	.02†	.11†	.06†	.28†	.24†	.05†
Mentoring Others	.23**	.13*	.15**	.00†	.06†	.11*	.10†	.13†	.06†

Note. T1 = Retesting on T1 after taking T1; T2 = Retesting on T2 after taking T2; T1/T2 or T2/T1 = Retesting on the Opposite Form.
Effect size = Cohen's *d*. NA indicates that the parameters could not be estimated due to a small samples size. OTS-CIV: N_{T1} = 194, N_{T2} = 146, $N_{\text{T1/T2 or T2/T1}}$ = 738; OTS-AD: N_{T1} = 200, N_{T2} = 166, $N_{\text{T1/T2 or T2/T1}}$ = 1,290; AECP: N_{T1} = 0, N_{T2} = 2, $N_{\text{T1/T2 or T2/T1}}$ = 12; USAFA: N_{T1} = 162, N_{T2} = 16, $N_{\text{T1/T2 or T2/T1}}$ = 804; ROTC: N_{T1} = 702, N_{T2} = 676, $N_{\text{T1/T2 or T2/T1}}$ = 5,842; ANG: N_{T1} = 126, N_{T2} = 82, $N_{\text{T1/T2 or T2/T1}}$ = 884; AFRES: N_{T1} = 38, N_{T2} = 40, $N_{\text{T1/T2 or T2/T1}}$ = 362.

APPENDIX T. Stratified Samples Subtest-Level Stability Analysis

OTS-CIV

SUBTEST	T1						T2					
	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.
Overall	2.64	.14	2.65	.13	.07*	.01	2.66	.14	2.66	.14	$.00^{\dagger}$	.00
Most Effective	2.60	.17	2.62	.17	.10**	.02	2.61	.17	2.62	.17	$.04^{\dagger}$	.01
Least Effective	2.69	.15	2.69	.14	$.01^{\dagger}$	.00	2.71	.15	2.70	.15	$.04^{\dagger}$	.01
Integrity	2.68	.23	2.71	.23	.13**	.03	2.52	.37	2.53	.35	$.03^{\dagger}$	.01
Leading Others	2.64	.18	2.67	.18	.17**	.03	2.73	.21	2.73	.21	$.02^{\dagger}$	.00
Decision Making	2.52	.28	2.51	.27	$.04^{\dagger}$	.01	2.60	.22	2.57	.23	.12**	.03
Communication Skills	2.66	.30	2.65	.29	$.01^{\dagger}$	.00	2.63	.23	2.64	.22	.07*	.02
Leading Innovation	2.77	.33	2.74	.33	.07*	.02	2.79	.41	2.75	.45	.10**	.04
Mentoring Others	2.66	.25	2.65	.26	$.02^{\dagger}$	.01	2.68	.20	2.68	.20	$.02^{\dagger}$	.00

*$p \le .05$; **$p \le .01$; $^{\dagger}p$ not significant.

Note. M_{early} = Mean of the Early Test Administration Group; SD_{early} = Standard Deviation of the Early Administration Group; M_{later} = Mean of the Later Test Administration Group; SD_{later} = Standard Deviation of the Later Test Administration Group; ES = Effect Size (Cohen's *d*); Abs. *M* Diff. = Absolute Mean Difference. N_{T1} = 1,800; N_{T2} = 1,600.

OTS-AD

SUBTEST	T1						T2					
	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.
Overall	2.71	.12	2.70	.12	.09*	.01	2.71	.12	2.69	.13	.10*	.01
Most Effective	2.67	.15	2.67	.16	$.04^{\dagger}$	.01	2.66	.16	2.65	.16	$.07^{\dagger}$	.01
Least Effective	2.74	.13	2.72	.13	.11**	.01	2.75	.13	2.74	.13	.10*	.01
Integrity	2.76	.21	2.76	.22	$.01^{\dagger}$	.00	2.56	.34	2.56	.35	$.00^{\dagger}$	.00
Leading Others	2.70	.17	2.70	.17	$.03^{\dagger}$	.00	2.75	.20	2.75	.20	$.01^{\dagger}$	.00
Decision Making	2.55	.25	2.52	.26	.12**	.03	2.60	.20	2.57	.21	.13**	.03
Communication Skills	2.71	.29	2.69	.29	$.06^{\dagger}$	.02	2.70	.19	2.68	.20	.14**	.03
Leading Innovation	2.80	.29	2.79	.30	$.05^{\dagger}$	.02	2.83	.37	2.85	.35	$.04^{\dagger}$	.01
Mentoring Others	2.75	.21	2.75	.22	$.01^{\dagger}$	.00	2.76	.17	2.76	.17	$.02^{\dagger}$	.00

*$p \le .05$; **$p \le .01$; $^{\dagger}p$ not significant.

Note. M_{early} = Mean of the Early Test Administration Group; SD_{early} = Standard Deviation of the Early Administration Group; M_{later} = Mean of the Later Test Administration Group; SD_{later} = Standard Deviation of the Later Test Administration Group; ES = Effect Size (Cohen's *d*); Abs. *M* Diff. = Absolute Mean Difference. N_{T1} = 1,400; N_{T2} = 1,200.

AECP

SUBTEST	T1						T2					
	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.
Overall	2.71	2.68	.08	.14	.25†	2.64	2.67	2.66	.12	.14	.13†	2.56
Most Effective	2.67	2.66	.12	.20	.05†	2.55	2.62	2.59	.14	.20	.13†	2.47
Least Effective	2.76	2.71	.11	.13	.42†	2.65	2.73	2.72	.11	.13	.08†	2.63
Integrity	2.72	2.75	.19	.20	.13†	2.53	2.53	2.58	.32	.27	.14†	2.21
Leading Others	2.72	2.65	.17	.22	.38†	2.55	2.71	2.68	.16	.22	.12†	2.54
Decision Making	2.54	2.52	.23	.22	.07†	2.31	2.55	2.58	.25	.20	.15†	2.30
Communication Skills	2.77	2.78	.24	.24	.06†	2.53	2.66	2.63	.16	.23	.16†	2.50
Leading Innovation	2.80	2.79	.23	.31	.04†	2.57	2.82	2.83	.38	.42	.04†	2.43
Mentoring Others	2.79	2.71	.17	.24	.38†	2.62	2.77	2.71	.18	.19	.34†	2.60

$^{\dagger}p$ not significant.

Note. M_{early} = Mean of the Early Test Administration Group; SD_{early} = Standard Deviation of the Early Administration Group; M_{later} = Mean of the Later Test Administration Group; SD_{later} = Standard Deviation of the Later Test Administration Group; ES = Effect Size (Cohen's *d*); Abs. *M* Diff. = Absolute Mean Difference. N_{T1} = 25; N_{T2} = 30.

USAFA

SUBTEST	T1						T2					
	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.
Overall	2.72	.10	2.68	.12	.30**	.03	2.71	.10	2.68	.13	.27**	.03
Most Effective	2.70	.14	2.66	.15	.22**	.03	2.68	.13	2.65	.16	.21**	.03
Least Effective	2.74	.12	2.71	.13	.27**	.03	2.75	.11	2.72	.14	.25**	.03
Integrity	2.78	.19	2.75	.21	.17**	.04	2.53	.33	2.48	.36	.16**	.06
Leading Others	2.70	.17	2.69	.17	.06†	.01	2.81	.17	2.80	.20	.08†	.01
Decision Making	2.61	.23	2.52	.25	.34**	.08	2.62	.19	2.57	.22	.24**	.05
Communication Skills	2.69	.28	2.66	.30	.12**	.03	2.68	.19	2.65	.21	.16**	.03
Leading Innovation	2.78	.30	2.77	.32	.03†	.01	2.83	.38	2.84	.38	.02†	.01
Mentoring Others	2.77	.19	2.74	.21	.14**	.03	2.75	.16	2.72	.18	.18**	.03

*$p \leq .05$; **$p \leq .01$; $^{\dagger}p$ not significant.

Note. M_{early} = Mean of the Early Test Administration Group; SD_{early} = Standard Deviation of the Early Administration Group; M_{later} = Mean of the Later Test Administration Group; SD_{later} = Standard Deviation of the Later Test Administration Group; ES = Effect Size (Cohen's *d*); Abs. *M* Diff. = Absolute Mean Difference. N_{T1} = 1,100; N_{T2} = 900.

ROTC

SUBTEST	T1						T2					
	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.
Overall	2.64	.13	2.63	.14	.06†	.01	2.63	.15	2.64	.14	.10**	.01
Most Effective	2.60	.17	2.59	.17	.03†	.01	2.57	.17	2.59	.17	.10**	.02
Least Effective	2.68	.15	2.67	.15	.06†	.01	2.68	.16	2.69	.14	.07*	.01
Integrity	2.70	.23	2.71	.23	.02†	.00	2.48	.37	2.48	.37	.00†	.00
Leading Others	2.65	.18	2.65	.19	.02†	.00	2.70	.23	2.72	.21	.10**	.02
Decision Making	2.49	.28	2.47	.27	.08†	.02	2.55	.22	2.56	.22	.02†	.00
Communication Skills	2.61	.32	2.59	.32	.04†	.01	2.59	.22	2.61	.21	.07*	.02
Leading Innovation	2.75	.33	2.74	.35	.05†	.02	2.78	.42	2.81	.40	.06†	.03
Mentoring Others	2.65	.26	2.64	.27	.03†	.01	2.65	.21	2.67	.20	.07*	.01

$*p \leq .05$; $**p \leq .01$; $^{\dagger}p$ not significant.

Note. M_{early} = Mean of the Early Test Administration Group; SD_{early} = Standard Deviation of the Early Administration Group; M_{later} = Mean of the Later Test Administration Group; SD_{later} = Standard Deviation of the Later Test Administration Group; ES = Effect Size (Cohen's *d*); Abs. *M* Diff. = Absolute Mean Difference. N_{T1} = 1,900; N_{T2} = 1,800.

ANG

SUBTEST	T1						T2					
	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.
Overall	2.70	.11	2.69	.12	.08†	.01	2.70	.13	2.69	.14	.02†	.00
Most Effective	2.67	.15	2.66	.15	.10*	.01	2.65	.17	2.64	.17	.01†	.00
Least Effective	2.73	.13	2.72	.13	.04†	.00	2.75	.13	2.74	.15	.02†	.00
Integrity	2.75	.21	2.75	.22	.02†	.00	2.58	.34	2.57	.34	.03†	.01
Leading Others	2.69	.17	2.68	.18	.07†	.01	2.75	.20	2.76	.21	.06†	.01
Decision Making	2.57	.26	2.56	.25	.05†	.01	2.61	.20	2.59	.22	.08†	.02
Communication Skills	2.70	.28	2.69	.27	.05†	.01	2.67	.21	2.66	.21	.03†	.01
Leading Innovation	2.78	.30	2.76	.34	.04†	.01	2.82	.38	2.80	.40	.04†	.02
Mentoring Others	2.74	.21	2.73	.22	.03†	.01	2.74	.18	2.75	.20	.02†	.00

$*p \leq .05$; $^{\dagger}p$ not significant.

Note. M_{early} = Mean of the Early Test Administration Group; SD_{early} = Standard Deviation of the Early Administration Group; M_{later} = Mean of the Later Test Administration Group; SD_{later} = Standard Deviation of the Later Test Administration Group; ES = Effect Size (Cohen's *d*); Abs. *M* Diff. = Absolute Mean Difference. N_{T1} = 1,000; N_{T2} = 800.

AFRES

SUBTEST	T1						T2					
	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.	M_{early}	SD_{early}	M_{later}	SD_{later}	ES	Abs. *M* Diff.
Overall	2.69	.13	2.69	.12	.00†	.00	2.68	.14	2.70	.12	.13†	.02
Most Effective	2.65	.17	2.66	.15	.06†	.01	2.63	.16	2.65	.16	.11†	.02
Least Effective	2.72	.14	2.71	.13	.07†	.01	2.73	.15	2.75	.13	.10†	.01
Integrity	2.72	.22	2.73	.22	.03†	.01	2.57	.34	2.56	.36	.03†	.01
Leading Others	2.66	.18	2.69	.17	.16*	.03	2.72	.21	2.75	.19	.15†	.03
Decision Making	2.58	.24	2.55	.25	.13†	.03	2.61	.21	2.61	.20	.00†	.00
Communication Skills	2.69	.29	2.71	.26	.06†	.02	2.66	.21	2.67	.19	.07†	.01
Leading Innovation	2.80	.29	2.75	.31	.19*	.06	2.79	.42	2.80	.42	.03†	.01
Mentoring Others	2.73	.22	2.72	.23	.07†	.01	2.73	.19	2.75	.17	.12†	.02

*$p \leq .05$; $^{\dagger}p$ not significant.

Note. M_{early} = Mean of the Early Test Administration Group; SD_{early} = Standard Deviation of the Early Administration Group; M_{later} = Mean of the Later Test Administration Group; SD_{later} = Standard Deviation of the Later Test Administration Group; ES = Effect Size (Cohen's *d*); Abs. *M* Diff. = Absolute Mean Difference. $N_{\text{T1}} = 300$; $N_{\text{T2}} = 250$.

APPENDIX U. Additional Analyses

A couple of exploratory analyses were conducted to examine the attributes of the SJT. Specifically, Generalizability Theory (G-Theory) was used as a technique to estimate the SJT internal structure (i.e., reliability) and stepwise regression was used to examine how much variance in SJT is accounted for by demographic variables.

Generalizability Theory

Generalizability Theory (G-Theory) allows for partitioning variance to various sources (Jackson et al., 2017). In the case of the SJT, these sources of variance were: test-takers, dimensions (i.e., competencies), situations (i.e., situations), items (i.e., ME/LE), and their interactions. Because the competencies were not balanced and potentially not well defined, competencies were excluded as a source of potential variance. This G-Theory application was tested for random samples of 5,000 test-takers. The models did not converge well and the proportions of variance accounted for among the different samples was not consistent. Therefore, it was determined that the G-Theory results were not robust. It is recommended that this approach be attempted again with stronger competency definitions.

Stepwise Regression

As a way of exploring whether or not demographics played role in the SJT responses, we performed stepwise regression analyses using demographic variables (e.g., age, sex, race) as predictors and the SJT subtest scores (Overall, ME, and LE) as outcomes. Stepwise regression is an atheoretical procedure and is therefore not usually recommended when *a priori* theory for the multivariate relationships is known. However, there was no underlying theory for the association among the demographic variables and the SJT responses, so we decided to use both forward and backward stepwise regression analysis. Note that these analyses were not performed for the stratified samples. The total amount of variance accounted for by the demographic variables ranged from 5.3% for LE to 6.2% for ME to 7.8% for Overall. Notably, the demographic variables accounted for more variance in the ME scores than in the LE scores, which is not surprising given that the ME scores tended to correlate higher with cognitive ability than did the LE scores. The Black race dummy-coded variable accounted for the largest amount of variance of all of the demographic variables in each of the regressions. It was followed by age, Asian race, Hispanic White ethnicity, American Indian/Alaskan Native race, and Native Hawaiian/Pacific Islander race in each of the three regressions. Sex also explained a significant proportion of variance for ME and LE. The contributions of individual demographic variables alone were relatively small by comparison as evidenced by rather small changes in R^2 (see table below).

Overall SJT Score				ME SJT Score				LE SJT Score			
Step	Variable	R^2	ΔR^2	Step	Variable	R^2	ΔR^2	Step	Variable	R^2	ΔR^2
1	Black	4.10%		1	Black	3.60%		1	Black	2.40%	
2	Age	5.90%	1.80%	2	Age	4.80%	1.2%	2	Age	3.70%	1.3%
3	Asian	7.10%	1.20%	3	Asian	5.60%	.80%	3	Asian	4.80%	1.1%
4	HispW	7.50%	.40%	4	HispW	5.90%	.30%	4	HispW	5.10%	.30%
5	AIAN	7.70%	.20%	5	AIAN	6.00%	.10%	5	AIAN	5.20%	.10%
6	NHPI	7.80%	.10%	6	NHPI	6.10%	.10%	6	NHPI	5.30%	.10%
				7	Sex	6.20%	.10%	7	Sex	5.30%	.00%

SYMBOLS, ABBREVIATIONS AND ACRONYMS

A	An abbreviation used to denote Asian race
AA	African American
ABM	Air Battle Manager
AD	Active Duty
AD-Other	Active Duty Other
AECP	Airman Education and Commissioning Program
AFB	Air Force Base
AFOQT	Air Force Officer Qualifying Test
AFPC/DSYX	Air Force Personnel Center Strategic Research and Assessments Branch
AFRES	Air Force Reserve
AI	Aviation Information
AIAN	American Indian/Alaska Native
ANG	Air National Guard
AR	Arithmetic Reasoning
ASVAB	Armed Services Vocational Aptitude Battery
B	An abbreviation used to denote Black race
BC	Block Counting
BH	An abbreviation used to denote Black race and Hispanic ethnicity
BnH	An abbreviation used to denote Black race and non-Hispanic ethnicity
CIT	Critical Incidents Technique
CIV	Civilian
CIV-Other	Civilian Other
Cohen's *d*	Effect Size
Comm.	Communication Skills
CSO	Combat Systems Operator
CTT	Classical Test Theory
D&I	Diversity and Inclusion
Dec.	Decision Making
df	degrees of freedom
EFA	Exploratory Factor Analysis
ES	Effect Size
F	An abbreviation used to denote female
G-Theory	Generalizability Theory
IC	Instrument Comprehension
IER	Insufficient Effort Responding
Inn.	Leading Innovation
Integ.	Integrity
IST	Infoscitex
ITC	Item-Total Correlation
ITP	Implicit Trait Policy
LE	Least Effective
Lead.	Leading Others

M	A symbol used to denote the mean
M	An abbreviation used to denote male
MAR	Missing at Random
Max	Maximum
MCAR	Missing Completely at Random
ME	Most Effective
Ment.	Mentoring Others
Min	Minimum
MK	Math Knowledge
MNAR	Missing not at Random
MPA	Military Performance Average
N	A symbol used to denote number of data points (i.e., responses, test-takers) in a sample
NA	Not Available
NHPI	Native Hawaiian/Other Pacific Islander
O-1	Air Force Second Lieutenant
O-2	Air Force First Lieutenant
O-3	Air Force Captain
OSM	Officer Suitability Measure
OTS	Officer Training School
OTS-AD	Officer Training School-Active Duty
OTS-CIV	Officer Training School-Civilian
p	Probability value (significance level)
PCA	Principal Components Analysis
PII	Personally Identifiable Information
PS	Physical Science
PWS	Performance Work Statement
r	Correlation
R^2	R-Squared (Proportion of Variance Explained)
RC	Reading Comprehension
Resp.	Response
RGL	Reading Grade Level
ROTC	Reserve Officer Training Corps
RPA	Remotely-Piloted Aircraft
RSS	Relative Standing Score
SD	A symbol used to denote the standard deviation
SDI-O	Self-Description Inventory-Officer
Sig.	Significance Level
SJT	Situational Judgment Test
SJT-SCO	Variable used to denote the SJT subtest score with criterion-invalid items removed
SME	Subject Matter Expert
SPARKID	A dummy identification number used to assist in identifying data without the use of PII
t	A symbol used to denote the t-statistic

T1	Air Force Officer Qualifying Test Form T Form 1
T2	Air Force Officer Qualifying Test Form T Form 2
TCO-TE	Test Control Officer-Test Evaluator
TR	Table Reading
URT	Undergraduate RPA Training
USAF	United States Air Force
USAFA	United States Air Force Academy
VA	Verbal Analogies
W	An abbreviation used to denote White race
WH	An abbreviation used to denote White race and Hispanic ethnicity
WK	Word Knowledge
WnH	An abbreviation used to denote White race and non-Hispanic ethnicity
α	Cronbach’s alpha
ω	McDonald’s omega