AFRL-RH-WP-TR-2019-0067

EVALUATION OF CRITERION-RELATED VALIDITY AND POTENTIAL ITEM EXPOSURE EFFECTS FOR THE WEIGHTED AIRMAN PROMOTION SYSTEM (WAPS)

Kevin M. Bradley
Jeffrey A. Dahlke
Rodney A. McCloy
Matthew C. Reeder
Martin Yu

Human Resources Research Organization (HumRRO)

September 2019
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2019-0067 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//
THOMAS R. CARRETTA
Work Unit Manager
Collaborative Interfaces and Teaming Branch
Warfighter Interfaces Division

//signature//
TIMOTHY S. WEBB
Chief, Collaborative Interfaces and Teaming Branch
Warfighter Interfaces Division

//signature//
LOUISE A. CARTER
Chief, Warfighter Interfaces Division
Airman Systems Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30-09-19	Interim	1 JAN 18 – 30 SEP 19

4. TITLE AND SUBTITLE
Evaluation of Criterion-Related Validity and Potential Item Exposure Effects for the Weighted Airman Promotion System (WAPS)

5a. CONTRACT NUMBER
FA8650-14-D-6500, 0007

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Kevin M. Bradley, Jeffrey A. Dahlke, Rodney A. McCloy, Matthew C. Reeder, and Martin Yu

5d. PROJECT NUMBER
5329

5e. TASK NUMBER
03

5f. WORK UNIT NUMBER
H0SA 532909TC

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 700
Alexandria, VA 22314-1578

8. PERFORMING ORGANIZATION REPORT NUMBER
2019 No. 093

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Material Command
Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Warfighter Interface Division
Collaborative Interfaces and Teaming Branch
Wright-Patterson AFB, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHCC

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2019-0067

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A; Approved for public release. 88ABW-2019-4979. Cleared 10/29/2019

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Weighted Airman Promotion System (WAPS) determines promotions to non-commissioned officer (NCO) ranks within the U.S. Air Force (USAF). The WAPS comprises a formula for weighting various components characterizing an Airman's readiness for promotion. Two standardized tests (multiple-choice format, each having 100 items) serve as WAPS components: (a) a Specialty Knowledge Test (SKT) – a measure of technical knowledge pertaining to the Air Force specialty (AFS; i.e., job) to which the individual belongs, and (b) the Promotion Fitness Exam (PFE) – a measure of general USAF knowledge covering topics such as history, customs, dress and appearance, resource management, and security. SKTs are specific to each AFS, but the PFE is given to all members of a given rank, regardless of AFS. This report provides the results of analyses conducted to address the criterion-related validity of the WAPS tests and the extent to which that validity might be reduced by repeated exposure of test content. The potential for differential item exposure effects by race/ethnicity and gender is also examined.

15. SUBJECT TERMS
Weighted Airman Promotion System, Performance Fitness Exam, Specialty Knowledge Test

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Thomas R. Carretta
Unclassified	Unclassified	Unclassified	SAR	127	19b. TELEPHONE NUMBER *(Include Area Code)* N/A

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Exhibits

EXECUTIVE SUMMARY

The Weighted Airman Promotion System (WAPS) determines promotions to non-commissioned officer (NCO) ranks within the U.S. Air Force. Two standardized tests – the Specialty Knowledge Test (SKT) and the Promotion Fitness Exam (PFE) – serve as WAPS components. Although these tests were developed to rigorous standards to ensure their content validity, the Air Force has sought additional evidence to ensure accord with best practice. Consequently, the Air Force commissioned the Human Resources Research Organization (HumRRO) to carry out the present project consisting of two primary tasks:

- Task 1: Evaluate the criterion-related validity of the WAPS tests relative to Airman performance. As part of these analyses, we also evaluated the operational PFE/SKT standards in terms of their ability to differentiate Airman performance.
- Task 2: Investigate the extent to which item exposure might (a) affect psychometric properties and lead to attenuated validity of the WAPS tests and (b) differentially affect examinees of various demographic groups.

Task 1 results provided little evidence that SKT and PFE scores predict Enlistment Performance Report (EPR) ratings. Although there were some instances of non-negligible outcome prediction, there was little evidence overall regarding the predictive efficacy of the SKT or PFE. Similarly weak relations were observed for the Armed Services Vocational Aptitude Battery (ASVAB), which has substantial evidence of its criterion-related validity for job performance. We believe the lack of supportive validity findings in this study may be attributable more to properties of the outcome measures (criteria) than to deficiencies associated with the SKT or PFE. Specifically, the EPR ratings (a) were highly restricted in terms of score variability and (b) included non-technical, non-duty relevant considerations that appear relatively distinct from aspects of performance likely to be predicted by SKT or PFE scores. Although the EPR ratings might provide value to the Air Force from an operational standpoint, these administrative ratings might not be well-suited for validation research compared to measures that better differentiate Airman performance and are more relevant to the focal predictor constructs (e.g., research-only performance rating scales, hands-on work sample performance measures). Consequently, we believe that the results of this study provide inconclusive evidence regarding the predictive validity of the SKT and PFE.

Task 2 was motivated by the fact that Airmen have the opportunity to complete the SKT or PFE on multiple occasions. In general, results from the item exposure analyses provided little evidence that exposure affects properties of the items or examinee test scores. We found little evidence that item exposure systematically affects demographic subgroup item-level performance or relations between Airman experience and item-level performance. In addition, there was little evidence that item-level psychometric properties (i.e., difficulty, discrimination, correlations between item-level scores and external variables) differed between first-time and repeat examinees, or that item exposure moderated these differences. Results from differential item functioning (DIF) analyses provided no systematic evidence that items functioned differently between first-time and repeat examinees, regardless of whether the items had been exposed. Given the general lack of supportive findings concerning the predictive validity of the PFE and SKT in Task 1, we did not conduct analyses to investigate whether item exposure attenuates the criterion-related validity of the test scores.

1.0 INTRODUCTION

The Weighted Airman Promotion System (WAPS) determines promotions to non-commissioned officer (NCO) ranks within the U.S. Air Force (USAF). The WAPS comprises a formula for weighting various components characterizing an Airman's readiness for promotion. Two standardized tests (multiple-choice format, each having 100 items) serve as WAPS components. The first is the Specialty Knowledge Test (SKT) – a measure of technical knowledge pertaining to the Air Force specialty (AFS; i.e., job) to which the individual belongs. The second is the Promotion Fitness Exam (PFE) – a measure of general USAF knowledge covering topics such as history, customs, dress and appearance, resource management, and security. SKTs are specific to each AFS, but the PFE is given to all members of a given rank, regardless of AFS.

> Prior to the fall of 2019, the WAPS was used for promotion to all NCO ranks—that is, E-5 (Staff Sergeant) through E-9 (Chief Master Sergeant). Senior NCOs (i.e., ranks E-7 [Master Sergeant] through E-9) were promoted via a two-phase process. Phase one involved administration of both the SKT and PFE for promotion to E-7 and administration of only the PFE for promotion to E-8 (Senior Master Sergeant) and E-9. Phase two consisted of an evaluation by a promotion board. On 4 February 2019, the Air Force announced a change to promotions for senior NCOs that eliminates the first phase of the process (i.e., WAPS testing; Losey, 2019). Phasing out of the WAPS tests will begin during the fall of 2019 with promotions to E-9.

The WAPS tests are developed to rigorous standards to ensure their content validity. The Air Force has sought additional validity evidence to ensure accord with industry best practice (EEOC, 1978; SIOP, 2018) – specifically, criterion-related validity evidence, which is obtained by demonstrating empirical relations with meaningful outcome measures. For the Air Force, the outcome (i.e., criterion) of primary interest is work performance. Thus, Air Force tasked the Human Resources Research Organization (HumRRO) with establishing the criterion-related validity of the WAPS tests with regard to criterion measures of Airman performance – specifically, scores on Enlisted Performance Reports (EPR).[1] The Air Force also sought evidence that could be used to help evaluate the effects of the 2016 policy change requiring a minimum score of 40 on both the SKT and PFE, and of 90 for the sum of the SKT and PFE scores.

A second area of inquiry involved a subject common to all high-stakes testing programs: concerns regarding test exposure and concomitant test compromise, which could lead to reduced test validity. Candidates for promotion who are not selected during their first year of eligibility can test again as long as they remain eligible for the targeted higher rank (a period of 5 to 8 years). Thus, Airmen have the opportunity to complete the SKT or PFE on multiple occasions. Given the current retesting policy, promotion candidates could complete the SKT for E-6 as many as four to seven times. This leads to the likelihood that candidates will see many of the same test items upon re-administration, because approximately 20-40% of the questions for any given SKT appear on the SKT for the following year. This rate of item exposure raises questions about the potential for test compromise for later test administrations provided to a given candidate. Therefore, the Air Force asked HumRRO to investigate the extent to which item exposure was leading to attenuated validity of the WAPS

[1] Airmen are also evaluated on off-duty performance in areas such as continuing education.

measures and the extent to which exposure might be differentially affecting members according to their demographic characteristics, such as sex and racial/ethnic category.

In sum, the Air Force contracted with HumRRO to conduct two primary tasks related to the psychometric fitness of the WAPS tests:

- Task 1: Analyze Archival Data to Evaluate Overall Predictive Criterion-Related Validity of WAPS Tests
- Task 2: Analyze Archival Data Based on Examinees' First-Time and Repeated Item Exposure

The second task originally comprised two subtasks:

- Subtask 2.1. Compare demographic differences in WAPS exam performance (and associated adverse impact analyses, where applicable) based on item exposure.
- Subtask 2.2. Compare criterion-related validity of exam scores as a function of item exposure.

As described in the following section for Task 1, analyses pertaining to the criterion-related validity of the WAPS tests did not provide evidence that scores from the SKT or PFE are predictive of the outcomes examined in this study. Because we did not find evidence for predictive validity at the test score level, HumRRO did not perform analyses investigating the effect of item exposure on validity, subsumed under Subtask 2.2. Thus, results are presented only for Subtask 2.1.

The Air Force provided data to HumRRO to support all analyses to be conducted to complete the two tasks. We conducted the work comprised by Subtasks 2.1 through a series of five analyses. We therefore present the Task 2 results by the five sets/types of analyses conducted.

This report provides results of the analyses conducted to address the criterion-related validity of the WAPS tests and the extent to which demographic differences in WAPS exam performance that is affected by repeated exposure of test content. We begin with a discussion of the criterion-related validity analyses conducted on the two WAPS tests. We then turn our attention to analyses conducted to evaluate the impact of test exposure.

2.0 TASK 1: ANALYZE ARCHIVAL DATA TO EVALUATE OVERALL PREDICTIVE CRITERION-RELATED VALIDITY OF WAPS TESTS

The purpose of Task 1 was to evaluate the statistical relation between the Weighted Airman Promotion System (WAPS) exam performance—School Knowledge Test (SKT) and Promotion Fitness Examination (PFE)—and future Enlisted Performance Report (EPR) ratings (i.e., ratings obtained upon promotion). Analyses for this task were conducted to provide evidence for the criterion-related validity of the WAPS tests. Evaluating the criterion-related validity of a selection or promotion procedure (e.g., test scores, interview ratings) requires demonstrating an empirical relation between scores on that procedure and one or more relevant outcomes (Society for Industrial and Organizational Psychology, 2018).

The following is an overview of the analyses conducted:

- For examinees promoted to E-5, E-6, or E-7 in a given career field, examine relations between (a) SKT and PFE scores and (b) EPR ratings upon promotion (e.g., for Airmen who tested on a 2011 E-5 SKT and were promoted in that promotion cycle, the relation of 2011 E-5 SKT scores to aggregate 2012-2017 EPR ratings).
- Apply corrections for range restriction, as applicable. Corrections were based on the full SKT/PFE score range among all candidates (i.e., those who were and were not promoted) in a given cycle.
- Examine the incremental validity of WAPS test scores as a predictor of future EPR ratings, beyond other WAPS factors considered in the same WAPS testing cycle was examined. For example, for Airmen who tested on a 2011 E-5 SKT and were promoted in that cycle, examine the relation of 2011 E-5 SKT scores to aggregate 2012-2017 EPR ratings, over and above 2011 EPR scores, decorations, Time in Grade [TIG], and Time in Service [TIS]) was examined.
  - Because TIG and TIS are being phased out from future inclusion in WAPS, we also conducted a separate set of incremental validity analyses limited to EPR scores and decorations.
- Compare EPR ratings among promoted Airmen who (a) met versus (b) did not meet the minimum SKT/PFE cut scores of 40 and combined SKT/PFE cut score of 90.
- Control for Armed Services Vocational Aptitude Battery (ASVAB) scores (Armed Forces Qualifying Test score or the applicable Mechanical, Administrative, General, or Electronic composite score for a given career field) when examining relations among WAPS factors and future EPR ratings.

2.1 Data Files

Two data files containing variables relevant to the Task 1 analyses were provided to HumRRO: a Promotion Board data file and a Times Tested data file.

2.1.1 Promotion Board Data File

The Promotion Board data file contained scores on all factors considered during the Airman promotion evaluation process. These factors are described in Table 1.

Scores for prior performance ratings, decorations, TIG, and TIS in the Promotion Board data file were based on points awarded during the evaluation process. The Promotion Board data file also contained Airman background and demographic information (e.g., Air Force Specialty [AFS], race, ethnicity, gender) and promotion status (i.e., whether selected for promotion in a given cycle or not, reason for not being promoted). Finally, this data file also contained annual supervisor ratings on Airman performance and promotability, which would serve as outcome variables in all criterion-related validation analyses described in subsequent sections of this chapter.

Table 1. Factors Considered During Airman Promotion Evaluation Process.

Factor	Description
Promotion Fitness Examination (PFE) scores	Annual grade-specific exam (E-5 to E-7) measuring general Air Force knowledge
Specialty Knowledge Test (SKT) scores	Annual grade-specific exam (E-5 to E-7) measuring career field knowledge
Prior Performance (EPR) Ratings	Airman's Enlisted Performance Report (EPR) points in the cycle prior to promotion to the next grade
Decorations	Points awarded for decorations, with more prestigious awards being allocated more points
Time in Grade (TIG)	Points awarded for each month in current grade up until the first day of the last month of the promotion cycle (excluded from WAPS effective cycle 17)
Time in Service (TIS)	Points awarded for each month in Total Active Federal Military Service until the last day of the last month of the promotion cycle (excluded from WAPS effective cycle 17)

2.1.2 Times Tested Data File

The Times Tested data file contained Airman scores on the Armed Forces Qualification Test (AFQT) and the Air Force aptitude composites. The AFQT is a weighted composite of the four math and verbal ASVAB subtests: (a) Arithmetic Reasoning (AR), (b) Mathematical Knowledge (MK), (c) Paragraph Comprehension (PC), and (d) Word Knowledge (WK). A Verbal (VE) composite is first formed by unit-weighting scores on PC and WK. The AFQT score is then computed as AR+MK+2VE (i.e., a sum of the three standardized subtest scores), and that score is then converted to a percentile metric for reporting and screening purposes. The Air Force uses four aptitude composites for AFSC qualification. These aptitude composites, collectively referred to as "MAGE composites," are computed as follows:[2]

- Mechanical (M): AR+AS+MC+2VE
- Administrative (A): MK+VE
- General (G): AR+VE
- Electronics (E): AR+EI+GS+MK

2.2 Data Preparation

Prior to conducting our analyses, we implemented several steps to clean and prepare the Promotion Board and Times Tested data files for analysis.

2.2.1 Promotion Board Data File

The original structure of the Promotion Board data file was such that each row represented a

[2] Automotive and Shop Information (AS), Mechanical Comprehension (MC), Electronics Information (EI), General Science (GS).

unique Airman-cycle combination. Specifically, for a given Airman, each row contained updated Promotion Board data for each cycle between 2011 and 2017 (e.g., updated performance and promotability ratings, updated points on each promotion board factor and promotion status). An important first step in preparing the Promotion Board data file was to filter the file so that relevant data applicable to each cycle where an Airman was promoted were retained. This involved identifying cycles during which an Airman was promoted to one of the focal grades and identifying that Airman's PFE and SKT scores (as well as scores on other promotion factors) used in the promotion decision to serve as predictor scores in subsequent validation analyses.[3]

Restructuring the Promotion Board data also involved aggregating the supervisor performance and promotability ratings from after the cycle in which the Airman was promoted through all subsequent cycles in which the Airman maintained the same promoted-to grade. This step ensured that the data structure followed a predictive validation design, where predictor test scores used for promotion decision making are evaluated against subsequent post-promotion outcomes (Van Iddekinge & Ployhart, 2008). For each Airman, post-promotion ratings were identified and aggregated by locating ratings in cycles after the promotion decision, leading up to the next promotion event, where applicable. Ratings were included only up through the next promotion event to ensure that aggregation would occur only within the grade the Airman recently attained. For example, if an Airman was promoted to E-5, we identified all ratings for that Airman once assigned to E-5 up through the cycle when they were promoted to E-6. Airmen were excluded from aggregation if they were (a) never promoted within the 2011-2017 timeframe, (b) promoted only in the final cycle examined, or (c) selected for promotion but never formally progressed to the promoted-to grade. Cases meeting any of these three criteria were excluded because they did not have post-promotion ratings against which to validate the promotion factor scores.

Three types of post-promotion performance and promotability ratings were aggregated for analysis, including (a) "Legacy" performance ratings from the EPR in operational use prior to 2015, (b) new "EVAL" performance ratings from the EPR implemented in 2015, and (c) forced-distribution promotability ratings from the EPR implemented in 2015. Based on consultation with the Air Force, we prepared the ratings data in the following manner. Ratings were first dichotomized prior to aggregation to indicate whether the Airman received the maximum scale point possible in each cycle (i.e., 0 = maximum scale point not awarded, 1 = maximum scale point awarded; see Table 2). Accordingly, the Legacy ratings were recoded as follows: 5 = 1, 1–4 = 0. The new ratings were recoded as L = 1, E/V/A = 0. Two coding schemes were applied to the promotability ratings, based on consultation with the Air Force: (a) PN/MP = 1, DC/NN/PR = 0; (b) PR/PN/MP = 1, DC/NN = 0. Values on the post-promotion ratings not listed above were treated as missing.

We examined several methods of aggregating the dichotomized outcomes. Each aggregation method resulted in a value in percentage metric, representing the percent of possible post-promotion ratings where the maximum possible rating was obtained. Accordingly, the aggregated ratings ranged from 0% (i.e., no post-promotion ratings for an Airman obtained the maximum value) to 100% (i.e., that all post-promotion ratings for an Airman were equal to the maximum possible rating). Aggregation was performed for each rating separately, due to concerns about lack of scale and construct equivalence. For each rating format, the following

[3] Promoted Airmen were identified using the Promotion Select Cycle Code variable (CLP).

aggregation schemes were applied: (a) aggregate over the first two within-grade post-promotion ratings, (b) aggregate over the first three within-grade post-promotion ratings, and (c) aggregate over all within-grade post-promotion ratings.

Table 2. Dichotomization of Post-Promotion Rating Systems

Post-Promotion Rating	**Rating Dichotomization**	
	0	**1**
Legacy (1-5)	1 to 4	5
EVAL	E: Met some but not all expectations V: Met all expectations A: Exceeded some, but not all expectations	L: Exceeded most, if not all expectations
Forced Distribution (Approach 1)	DC: Do Not Promote NN: Not Ready Now PR: Promote	PN: Promote Now MP: Must Promote
Forced Distribution (Approach 2)	DC: Do Not Promote NN: Not Ready Now	PR: Promote PN: Promote Now MP: Must Promote

In addition to restructuring the promotion data and aggregating the post-promotion ratings, several other steps were involved in preparing the Promotion Board data:

- *Recoding invalid or out-of-range values to missing*.
- *Creating final analysis versions of race and ethnicity variables*. Race and ethnicity were treated as separate variables for analysis purposes.
- *Creating final analysis versions of the PFE and SKT scores*. Scores were created from the Promotion SKT Score (CLR3) and PFE USAFSE Score (CLR1) variables. In the original Promotion Board data file, PFE scores were doubled for SKT-exempt Airmen. Accordingly, we first identified SKT-exempt Airmen using criteria provided by the Air Force.[4] Among those Airmen identified as SKT-exempt, we then divided their PFE score by two and ensured that their SKT value was set to missing.
- *Creating final analysis versions of the other promotion factor scores*. These included time in grade (CLV), time in service (CLR5), promotion EPR score (CMA), promotion decorations (CLR), and promotion board and total scores (CBV11 and CMA2). Any instances of observed promotion factor scores exceeding the score limits for each grade and cycle were reset to the maximum score based on information provided by the Air Force (WAPS Points Changes Excel file).

[4] SKT-exempt Airmen were those (a) seeking promotion to grades E5 to E7, (b) with a valid non-missing PFE score, and (c) with a missing value for SKT ID Taken (AHS63).

2.2.2 Times Tested Data File

Preparation of the Times Tested data file for Task 1 analyses was relatively straightforward. In particular, we reset any out-of-range AFQT and MAGE aptitude composite scores to missing before subsequently merging these scores into the aggregated promotion board data file using a common subject identifier variable provided by the Air Force in each data file.

Table 3 lists all predictor, outcome, and control variables examined in the Task 1 analyses.

Table 3. Predictors, Outcomes, and Control Variables

WAPS Predictors	Post-Promotion Ratings	Controls
Promotion Fitness Examination (PFE)	Legacy (1-5) Ratings: 2011-2015	Armed Forces Qualification Test (AFQT)
Specialty Knowledge Test (SKT)	EVAL Ratings: 2015-2017	USAF Mechanical Aptitude Composite Score
Prior Performance (EPR) Ratings	Forced Distribution EPR Ratings: 2015-2017	USAF Administrative Aptitude Composite Score
Decorations		USAF General Aptitude Composite Score
Time in Grade (TIG)		USAF Electronics Aptitude Composite Score
Time in Service (TIS)		

2.2.3 Analysis Samples

Eight representative AFSs were selected for analysis. All of them had large testing populations and no major changes in 2011-2016 test content. They were

- Air Traffic Controller (1C1X1),
- Aircrew Flight Equipment (1P0X1),
- Weather Technician (1W0X1),
- Material (Supply) Management (2S0X1),
- Aircraft Armament Systems (2W1X1),
- Fire Protection (3E7X1),
- Security Forces (3P0X1), and
- Aerospace Medical Services (4N0X1).

The original data files received included Airmen from ranks, cycles, or AFSs outside of those listed above. Thus, we created filter variables to include only Airmen who met these criteria for rank, year, and AFS. Analyses where results are reported by rank, cycle, or AFS included only Airmen who fell in these groups. For analyses that involved fitting Bayesian mixed models with AFS as a random factor, we included all AFSs represented in the original data to improve estimation of AFS-level variance. Similarly, all analyses included only promoted Airmen except for the Bayesian mixed models, again to increase the number of level-two observations to

stabilize estimation.[5] An unrestricted, or population, data file used for multivariate range restriction corrections (Lawley, 1943) was also constructed by including all Airmen eligible for promotion (i.e., those who were and were not promoted) from the ranks, cycles, and AFSs listed above. Multivariate range restriction occurs when the variance on variables of interest in the analyzed sample is restricted relative to the variance in the population due to selection on these variables or correlated variables. The method described by Lawley (1943) was used to correct the sample-estimated correlations for multivariate range restriction among all analyzed variables as an estimate of the population-level correlations.

Cleaning, restructuring, and aggregation associated with preparing the Promotion Board and Times Tested data files resulted in three data files used for Task 1 analyses. The first data file was the **Promoted Airman Predictor/Outcome** data file ("Promoted Airmen"). In this data file, rows represented all cycles in which an Airman *was promoted* to one of the ranks of E-5 through E-7 between 2011 and 2017. This data file included all relevant predictor scores (i.e., scores on each of the promotion factors, including PFE and SKT exam scores), outcomes (i.e., aggregated performance and promotability ratings), control variables merged in from the original Times Tested data file (e.g., AFQT, MAGE scores), and other relevant Airman information (e.g., demographic and background variables). This data file was used for the primary correlational and regression analyses conducted to validate the promotion factors against Airman outcomes.

The second data file created was the **Unrestricted** data file ("Population"). In this data file, rows represented all cycles in which an Airman *was eligible for promotion* to one of the ranks of E-5 through E-7 between 2011 and 2017. This data file thus reflects the entire pool of promotion-eligible Airmen for each cycle and each grade; that is, all Airmen who *were* and who *were not* chosen for promotion. This Population data file was used for correcting correlations and comparable statistics (e.g., model R^2 estimates) estimated on the Promoted Airman data file. This data file contained only predictor scores and not aggregated post-promotion ratings.

The third data file created was the **Eligible Airman Predictor/Outcome** data file ("Eligible Airmen"). The Promoted Airman data file allowed for an evaluation of the predictive validity of WAPS promotion factors with post-promotion ratings. However, by restricting the analyses to only promoted Airman, statistics estimated by rank, cycle, and/or AFS were often based on very small sample sizes and, relatedly, very low outcome variability. We created the Eligible Airmen data file to be comparable in structure and contents to the Promoted Airmen data file, except that it included aggregated outcomes for Airmen who were *not* promoted in addition to those who *were* promoted. Including promoted and non-promoted Airmen allowed us to examine relations between promotion factors and aggregated outcomes on a larger sample of Airmen, which afforded greater estimation stability than the Promoted Airman data file. This data file was used only for analyses where the "Promoted Airmen" data file had insufficient sample sizes to estimate AFS-level variance—specifically, for the Bayesian mixed models. One caveat of deriving conclusions from the Eligible Airman data file is that relations estimated for non-promoted Airmen should be viewed as estimated on qualitatively different outcomes (i.e., ratings collected within the same grade) than those estimated for promoted Airmen (i.e., ratings collected within the promoted-to grade).

[5] The mixed models explicitly recognize the nested nature of the USAF data, with Airmen nested within AFSs. Airmen constitute level one in the Bayesian multilevel estimation model; AFSs constitute level two.

Tables 4 and 5 provide frequencies and percentages for demographic and background characteristics within the Population data file (n = 157,577) and the Promoted Airmen data file (n = 27,806), respectively. Tables 6 and 7 provide frequencies and percentages for demographic and background characteristics among promotion-eligible Airmen and promoted Airmen by promotion cycle.

Table 4. Population: Frequencies and Percentages and Demographic and Background Variables (n = 157,577)

Variable	Value	*n*	*%*
AFSC	1C1X1	8,902	6
	1P0X1	6,669	4
	1W0X1	6,684	4
	2S0X1	20,076	13
	2W1X1	19,259	12
	3E7X1	9,062	6
	3P0X1	70,881	45
	4N0X1	16,044	10
Promotion Cycle (Year)	2011	21,348	14
	2012	23,003	15
	2013	22,961	15
	2014	23,582	15
	2015	24,461	16
	2016	23,184	15
	2017	19,038	12
Promotion Grade	E5	75,905	48
	E6	54,175	34
	E7	27,497	17
Selected for Promotion	Not Selected	120,945	77
	Selected	36,632	23
Race	American Indian/Native Alaskan	1,142	1
	Asian	4,207	3
	Black/African American	32,796	21
	Native Hawaiian/Pacific Islander	2,672	2
	White	103,853	66
	More than One Race	5,717	4
Ethnicity	Hispanic or Latino	8,408	5
	Not Hispanic or Latino	149,079	95
Gender	Female	33,836	21
	Male	123,741	79

Note. AFSCs listed above were designated as focal AFSCs for analysis for this project. There are missing data for 7,190 Airmen on Race and 90 Airmen on Ethnicity.

Table 5. Promoted Airmen: Frequencies and Percentages for Demographic and Background Variables (n = 27,806)

Variable	Value	*n*	*%*
AFSC	1C1X1	1,931	7
	1P0X1	1,233	4
	1W0X1	1,424	5
	2S0X1	3,444	12
	2W1X1	3,304	12
	3E7X1	1,507	5
	3P0X1	12,016	43
	4N0X1	2,945	11
Promotion Cycle (Year)	2011	4,214	15
	2012	4,621	17
	2013	3,423	12
	2014	2,306	8
	2015	4,345	16
	2016	5,659	20
	2017	3,236	12
Promotion Grade	E5	16,459	59
	E6	7,332	26
	E7	4,013	14
Race	American Indian/Native Alaskan	193	1
	Asian	737	3
	Black/African American	4,942	18
	Native Hawaiian/Pacific Islander	452	2
	White	18,941	69
	More than One Race	994	4
Ethnicity	Hispanic or Latino	1,440	5
	Not Hispanic or Latino	26,010	95
Gender	Female	5,838	21
	Male	21,632	79

Note. AFSCs listed above were designated as focal AFSCs for analysis for this project. There are missing data for 2 Airmen on AFSC, Promotion Cycle, and Promotion Grade; 1,547 Airmen on Race; 356 Airmen on Ethnicity; and 336 on Gender.

Table 6. Population: Frequencies and Percentages for Demographic and Background Variables by Cycle

		2011		2012		2013		2014		2015		2016		2017	
Variable	**Value**	***n***	***%***	***n***	***%***	***n***	***%***	***n***	***%***	***n***	***%***	***n***	***%***	***n***	***%***
AFSC	1C1X1	1,062	5	1,226	5	1,274	6	1,391	6	1,464	6	1,388	6	1,097	6
	1P0X1	973	5	997	4	964	4	1,004	4	1,027	4	929	4	775	4
	1W0X1	819	4	900	4	963	4	1,016	4	1,087	4	1,049	5	850	4
	2S0X1	3,221	15	3,215	14	2,970	13	2,957	13	2,978	12	2,693	12	2,042	11
	2W1X1	2,746	13	2,852	12	2,738	12	2,809	12	3,054	12	2,842	12	2,218	12
	3E7X1	1,283	6	1,260	5	1,220	5	1,327	6	1,424	6	1,392	6	1,156	6
	3P0X1	9,183	43	10,310	45	10,431	45	10,640	45	10,971	45	10,560	46	8,786	46
	4N0X1	2,061	10	2,243	10	2,401	10	2,438	10	2,456	10	2,331	10	2,114	11
Promotion Grade	E5	9,337	44	10,585	46	10,725	47	11,069	47	12,096	49	11,782	51	10,311	54
	E6	7,869	37	8,520	37	8,380	37	8,315	35	8,226	34	7,589	33	5,276	28
	E7	4,142	19	3,898	17	3,856	17	4,198	18	4,139	17	3,813	16	3,451	18
Selected for Promotion	Not Selected	15,675	73	17,024	74	18,412	80	20,310	86	18,995	78	17,153	74	13,376	70
	Selected	5,673	27	5,979	26	4,549	20	3,272	14	5,466	22	6,031	26	5,662	30
Race	American Indian/Native Alaskan	179	1	176	1	162	1	168	1	168	1	159	1	130	1
	Asian	517	2	597	3	597	3	601	3	644	3	656	3	595	3
	Black/African American	4,669	22	5,066	22	4,916	21	4,870	21	4,877	20	4,594	20	3,804	20
	Native Hawaiian/Pacific Islander	341	2	393	2	413	2	426	2	417	2	380	2	302	2
	White	13,827	65	14,899	65	15,100	66	15,628	66	16,318	67	15,455	67	12,626	66
	More than One Race	578	3	704	3	729	3	821	3	1,000	4	1,009	4	876	5
Ethnicity	Hispanic or Latino	1,426	7	1,370	6	1,270	6	1,249	5	1,194	5	1,068	5	831	4
	Not Hispanic or Latino	19,915	93	21,622	94	21,677	94	22,317	95	23,253	95	22,099	95	18,196	96
Gender	Female	4,854	23	5,227	23	5,084	22	5,050	21	5,070	21	4,636	20	3,915	21
	Male	16,494	77	17,776	77	17,877	78	18,532	79	19,391	79	18,548	80	15,123	79

Note. AFSCs listed above were designated as focal AFSCs for analysis for this project.

Table 7. Promoted Airmen: Frequencies and Percentages for Demographic and Background Variables by Cycle

		2011		2012		2013		2014		2015		2016		2017	
Variable	**Value**	***n***	***%***	***n***	***%***	***n***	***%***	***n***	***%***	***n***	***%***	***n***	***%***	***n***	***%***
AFSC	1C1X1	258	6	328	7	257	8	197	9	295	7	411	7	185	6
	1P0X1	207	5	189	4	158	5	90	4	208	5	243	4	138	4
	1W0X1	161	4	197	4	175	5	191	8	226	5	355	6	119	4
	2S0X1	634	15	602	13	391	11	236	10	537	12	662	12	382	12
	2W1X1	545	13	559	12	385	11	217	9	516	12	696	12	386	12
	3E7X1	257	6	264	6	170	5	107	5	213	5	302	5	194	6
	3P0X1	1,763	42	2,018	44	1,515	44	878	38	1,907	44	2,437	43	1,498	46
	4N0X1	389	9	464	10	372	11	390	17	443	10	553	10	334	10
Promotion Grade	E5	2,284	54	2,673	58	2,151	63	1,236	54	2,474	57	3,660	65	1,981	61
	E6	1,342	32	1,349	29	776	23	590	26	1,075	25	1,291	23	909	28
	E7	588	14	599	13	496	14	480	21	796	18	708	13	346	11
Race	American Indian/Native Alaskan	36	1	35	1	27	1	7	0	34	1	33	1	21	1
	Asian	103	2	120	3	81	2	67	3	112	3	161	3	93	3
	Black/African American	797	19	885	19	632	19	449	20	708	17	898	16	573	18
	Native Hawaiian/Pacific Islander	80	2	65	1	61	2	32	1	74	2	87	2	53	2
	White	2,841	67	3,114	68	2,312	69	1,491	67	2,975	70	3,982	71	2,226	69
	More than One Race	119	3	134	3	104	3	82	4	164	4	245	4	146	5
Ethnicity	Hispanic or Latino	256	6	283	6	186	6	133	6	209	5	241	4	132	4
	Not Hispanic or Latino	3,957	94	4,303	94	3,181	94	2,093	94	4,048	95	5,353	96	3,075	96
Gender	Female	898	21	1,003	22	698	21	546	25	867	20	1,170	21	656	20
	Male	3,316	79	3,583	78	2,676	79	1,681	75	3,393	80	4,429	79	2,554	80

Note. AFSCs listed above were designated as focal AFSCs for analysis for this project.

2.2.4 Relate WAPS Exam Performance (SKT/PFE) to Future Enlisted Performance Report (EPR) Ratings

Descriptive statistics for the SKT and PFE scores among promoted Airmen (for the eight targeted AFSs only) are shown in Table 8. There was a good deal of variability in the average SKT and PFE scores across promoted-to AFS, cycle, and grade.

Table 8. Descriptive Statistics on SKT and PFE Scores for Promoted Airmen (Eight Targeted AFSs)

	SKT			PFE		
	n	**Mean**	**SD**	***n***	**Mean**	**SD**
Overall	27,444	64.51	9.62	27,804	63.54	8.48
Grade						
E5	16,371	63.97	8.84	16,459	62.82	7.69
E6	7,121	63.38	10.60	7,332	65.36	9.12
E7	3,952	68.82	9.72	4,013	63.16	9.73
Cycle						
11	4,214	57.15	9.20	4,214	56.81	8.08
12	4,586	57.77	8.64	4,621	61.32	8.08
13	3,374	63.43	8.21	3,423	64.83	8.01
14	2,227	70.18	7.92	2,306	64.21	7.64
15	4,248	67.56	7.38	4,345	63.56	7.09
16	5,593	68.86	7.97	5,659	67.36	7.67
17	3,202	69.42	6.93	3,236	66.94	7.35
AFSC						
1C1X1	1,912	68.46	8.28	1,931	68.05	8.30
1P0X1	1,215	58.41	11.61	1,233	61.53	8.36
1W0X1	1,420	63.80	8.34	1,424	66.71	7.92
2S0X1	3,365	64.41	10.33	3,444	61.41	8.92
2W1X1	3,272	64.32	11.16	3,304	62.27	8.33
3E7X1	1,499	64.01	8.52	1,507	65.28	8.51
3P0X1	11,868	64.90	9.04	12,016	63.11	8.19
4N0X1	2,893	63.81	8.75	2,945	64.70	7.85

Tables 9.a, 9.b, and 9.c display the descriptive statistics for the post-promotion ratings aggregated by the first two, first three, and all ratings, respectively. There was some variability in the average ratings across promoted-to AFS, cycle, and grade, but Airmen in higher grades tended to be rated higher, on average.

Table 9.a. Descriptive Statistics on Post-Promotion Ratings by Format for Promoted Airmen (8 Targeted AFSs, Aggregated First Two Post-Promotion Ratings)

	Legacy			New			Forced Distribution Approach 1			Forced Distribution Approach 2		
	n	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***	**Mean**	**SD**
Overall	12,521	.91	.23	8,397	.59	.41	4,483	.04	.14	4,483	.92	.20
Grade												
E5	7,120	.89	.25	5,080	.51	.41	2,894	.05	.15	2,894	.91	.21
E6	3,700	.92	.22	2,052	.67	.39	1,589	.02	.11	1,589	.94	.19
E7	1,701	.96	.16	1,265	.81	.31	0			0		
Cycle												
11	3,820	.91	.23	0			895	.03	.12	895	.91	.22
12	4,177	.91	.22	0			1,267	.04	.13	1,267	.90	.22
13	3,124	.90	.24	0			1,346	.04	.15	1,346	.92	.20
14	1,400	.90	.24	3	.50	.00	930	.04	.15	930	.95	.15
15	0			3,904	.55	.41	44	.03	.13	44	.91	.25
16	0			4,490	.63	.40	1	.00		1	.50	
17	0			0			0			0		
AFSC												
1C1X1	871	.92	.21	514	.62	.40	214	.02	.09	214	.93	.20
1P0X1	558	.90	.25	386	.55	.42	215	.01	.08	215	.91	.22
1W0X1	627	.89	.26	487	.47	.43	200	.03	.12	200	.92	.21
2S0X1	1,634	.91	.23	1,049	.62	.41	541	.04	.16	541	.91	.22
2W1X1	1,528	.94	.18	1,023	.61	.40	591	.05	.16	591	.92	.20
3E7X1	705	.87	.27	442	.58	.42	258	.03	.13	258	.90	.23
3P0X1	5,240	.91	.23	3,652	.59	.40	1,967	.04	.14	1,967	.92	.19
4N0X1	1,358	.89	.25	844	.61	.41	497	.04	.15	497	.93	.19

Note. The results presented were for the dichotomized ratings, where 1 = "Airmen earned the maximum possible rating."

Table 9.b. Descriptive Statistics on Post-Promotion Ratings by Format for Promoted Airmen (Eight Targeted AFSs Only, Aggregated First Three Post-Promotion Ratings)

	Legacy			New			Forced Distribution Approach 1			Forced Distribution Approach 2		
	n	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***	**Mean**	**SD**
Overall	8,508	.90	.21	6,079	.60	.41	2,819	.04	.12	2,819	.93	.16
Grade												
E5	4,720	.88	.22	3,847	.51	.41	1,765	.05	.13	1,765	.92	.16
E6	2,583	.91	.20	1,409	.72	.37	1,054	.02	.08	1,054	.94	.15
E7	1,205	.94	.16	823	.79	.33	0			0		
Cycle												
11	3,326	.90	.21	0			687	.03	.11	687	.93	.15
12	3,329	.90	.21	0			1,038	.04	.12	1,038	.92	.16
13	1,834	.89	.22	0			1,071	.04	.13	1,071	.93	.15
14	19	.82	.28	637	.47	.41	22	.02	.07	22	.94	.13
15	0			963	.56	.41	1	.00		1	1.00	
16	0			4,479	.63	.40	0			0		
17	0			0			0			0		
AFSC												
1C1X1	554	.92	.19	411	.61	.41	98	.02	.07	98	.96	.11
1P0X1	399	.88	.23	257	.53	.44	136	.01	.07	136	.93	.15
1W0X1	382	.87	.24	364	.46	.43	105	.03	.10	105	.92	.16
2S0X1	1,169	.90	.21	710	.63	.40	359	.04	.13	359	.93	.16
2W1X1	1,146	.93	.17	713	.63	.39	406	.05	.13	406	.93	.16
3E7X1	489	.87	.24	329	.57	.42	184	.02	.10	184	.90	.19
3P0X1	3,508	.90	.21	2,649	.61	.40	1,274	.04	.13	1,274	.93	.15
4N0X1	861	.88	.23	646	.60	.40	257	.03	.11	257	.93	.15

Note. The results presented were for the dichotomized ratings, where 1 = "Airmen earned the maximum possible rating."

Table 9.c. Descriptive Statistics on Post-Promotion Ratings by Format for Promoted Airmen (Eight Targeted AFSs Only, Aggregated All Post-Promotion Ratings)

	Legacy			New			Forced Distribution Approach 1			Forced Distribution Approach 2		
	n	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***	**Mean**	**SD**
Overall	14,455	.89	.24	21,678	.57	.43	8,369	.06	.20	8,369	.94	.18
Grade												
E5	8,308	.87	.26	12,547	.49	.43	5,612	.06	.22	5,612	.94	.19
E6	4,035	.90	.22	5,464	.64	.41	2,756	.04	.17	2,756	.95	.18
E7	2,112	.95	.17	3,667	.73	.38	1	.00		1	1.00	
Cycle												
11	4,147	.88	.23	1,771	.51	.42	1,208	.05	.18	1,208	.93	.18
12	4,578	.89	.23	2,402	.51	.42	1,789	.07	.22	1,789	.93	.18
13	3,405	.89	.25	2,231	.49	.41	1,652	.06	.19	1,652	.94	.17
14	2,293	.90	.26	2,062	.51	.41	1,242	.06	.20	1,242	.94	.19
15	28	.82	.39	4,328	.56	.39	2,405	.05	.21	2,405	.96	.19
16	1	1.00		5,652	.62	.42	61	.02	.13	61	.84	.36
17	3	.67	.58	3,232	.63	.48	12	.17	.39	12	.83	.39
AFSC												
1C1X1	1,034	.91	.22	1,422	.58	.43	420	.03	.15	420	.94	.20
1P0X1	638	.87	.26	985	.52	.43	391	.02	.12	391	.93	.19
1W0X1	721	.88	.25	1,158	.45	.43	397	.05	.21	397	.95	.18
2S0X1	1,842	.89	.23	2,652	.58	.42	1,022	.07	.23	1,022	.94	.19
2W1X1	1,697	.92	.20	2,718	.61	.42	1,066	.07	.22	1,066	.95	.17
3E7X1	793	.85	.27	1,195	.54	.44	452	.05	.20	452	.94	.19
3P0X1	6,124	.89	.24	9,214	.57	.42	3,666	.06	.21	3,666	.94	.19
4N0X1	1,606	.87	.27	2,334	.58	.43	955	.05	.19	955	.95	.18

Note. The results presented were for the dichotomized ratings, where 1 = "Airmen earned the maximum possible rating."

There were also considerable differences between the different EPR rating systems. In terms of the variability of EPR ratings, there appears to be low variability across the different formats and methods of aggregating post-promotion ratings. However, the new EVAL EPR ratings demonstrated more variability than did the legacy or forced distribution promotability rating. In the legacy system, Airmen were given the maximum rating over 90% of the time, indicating that raters did not make effective use of the full range of the scale. Conversely, raters utilized a wider range of the scale with the new EVAL EPR ratings, as Airmen were given the maximum rating about 60% of the time.

For the forced distribution promotability ratings, Airmen were given a rating of *Promote Now* or *Must Promote* about 5% of the time (aggregation approach 1). However, if a rating of *Promote* is also included in this aggregation (approach 2), then Airmen were found to have been given one of these three ratings about 90% of the time. In other words, most Airmen received the middle rating of *Promote*. This led to a rather ironic result: the variability in the forced distribution ratings was less than that in either the legacy or EVAL ratings. Forced distribution rating scale typically are employed to increase differentiation (i.e., variability) in operational performance ratings.

Correlations between the various aggregated future EPR ratings and the SKT and PFE are shown in Figure 1, corrected for multivariate range restriction (In Appendix A, Exhibit A1 shows these correlations without the correction). There is variability in the effectiveness of the SKT and PFE to predict post-promotion ratings at the promoted-to AFS, cycle, and grade levels, but due to low pairwise sample sizes when broken down by AFS, cycle, and/or grade, accurate conclusions cannot be drawn about differences between more specific groups. Overall, the correlations are weak at most. Specifically, the SKT and PFE have, on average, a correlation with legacy EPR ratings of about .10, and weaker or negligible correlations with the EVAL or forced distribution ratings. Additionally, the PFE tends to be a marginally better predictor than the SKT.

The joint effectiveness of the PFE and SKT for predicting future EPR ratings was examined using multilevel logistic regression. Due to issues with model convergence using a classical multilevel logistic model, the models were instead fitted using the framework of a Bayesian multilevel logistic regression in Stan (Stan Development Team, 2016). To maximize sample size, these models (a) were fitted using data from all available Airmen, regardless of whether they were promoted in a given cycle, and (b) included all AFSs in the data (i.e., including non-focal AFSs). Separate models were fitted for each grade and for each post-promotion EPR rating type and aggregation. The multilevel aspects of each model were specified as follows:

- AFS and cycle as random effects
- Predictors (SKT and PFE) as fixed effects and as random effects over AFS only

In this Bayesian modeling framework, the prior distributions of the parameters were defined as taking the form of a *t*-distribution (e.g., Gelman, Jaulin, Pittau, & Su, 2008). Estimation was carried out using a Markov chain Monte Carlo (MCMC) sampling algorithm with two chains and 100 iterations per chain. R^2 model fit values were calculated by correlating the fitted values from each model with the observed EPR ratings.

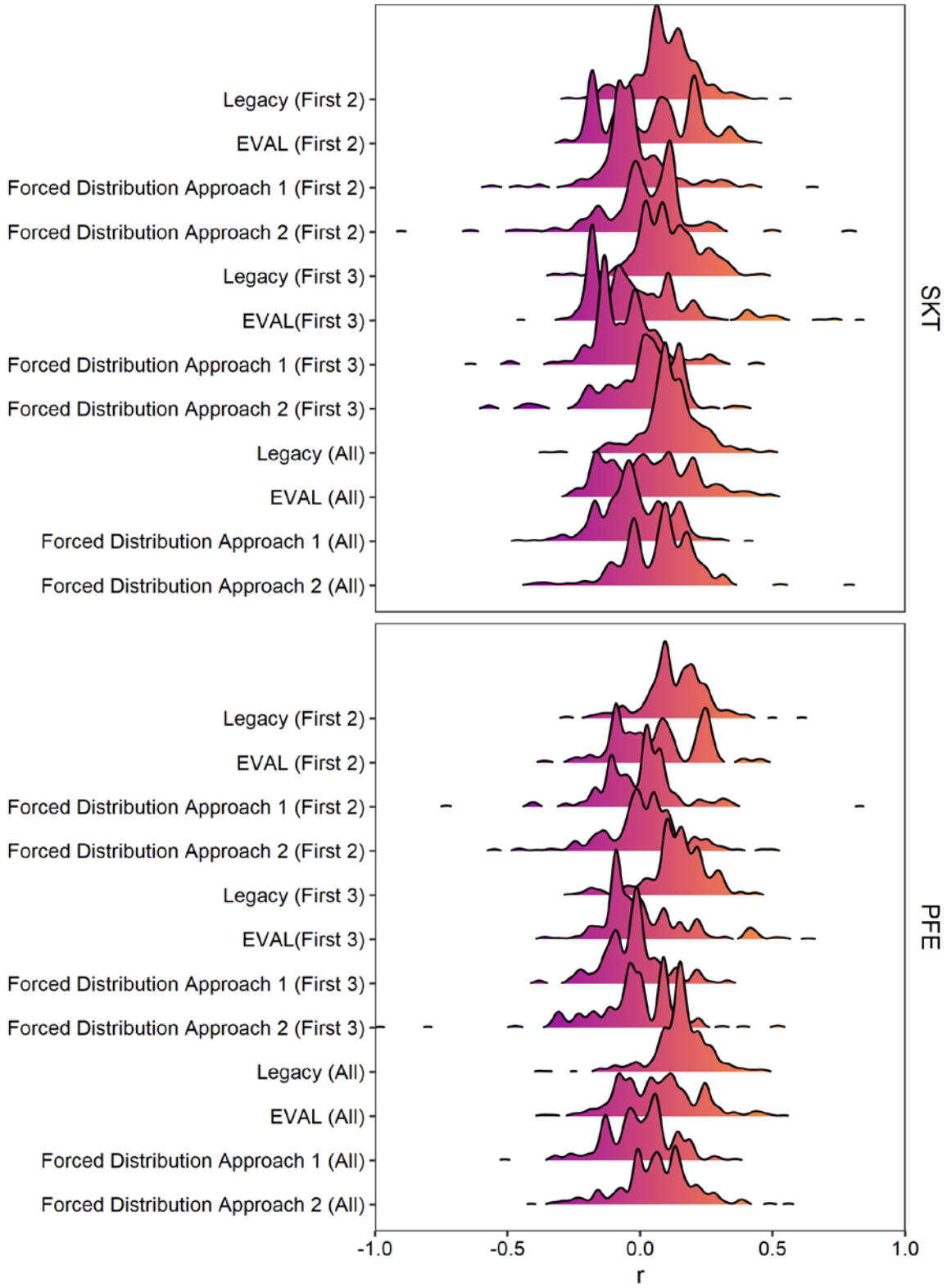
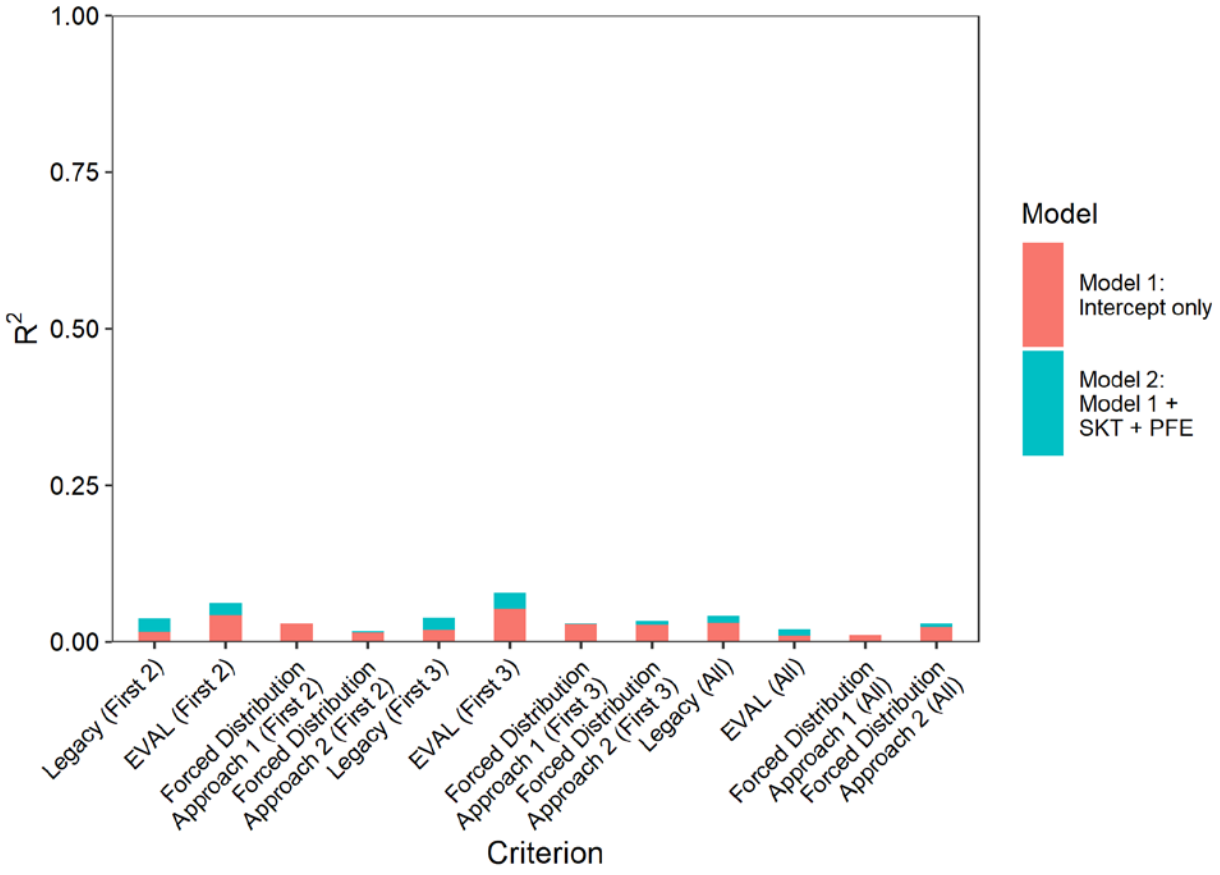


Figure 1. Distributions of predictor-outcome correlations across all AFSs, cycles, and grades, corrected for multivariate range restriction.

R^2 values for an intercept-only model and an intercept + SKT + PFE model are given in Figure 2. Overall, the SKT and PFE together provided little in terms of predicting future EPR ratings. Over the intercept-only model, they accounted for a few percent of the variance in the first two and first three post-promotion legacy or EVAL EPR ratings (R^2 between .01 and .04) and then basically did not add any additional prediction when all EPR ratings were aggregated. Furthermore, they accounted for basically zero variance in the forced distribution EPR ratings. However, we recommend that care be taken when interpreting these results. Because of the low variability in the EPR ratings as described previously, it is as yet unclear from these data whether it is the SKT and PFE that offer little prediction for an Airman's actual performance, or if it is the EPR ratings that offer little in meaningfully representing an Airman's actual performance. The latter case is highly likely given the low validities found when examining the other factors, including ASVAB composites, as presented in the following sections.



Note. Model 1 includes only the intercept term, and Model 2 includes SKT and PFE as additional predictors. The bar segment for Model 2 represents its increment in R^2 over Model 1. Models were fitted with AFS and cycle as random effects, and predictors as fixed effects and random effects over AFS only. Resulting R^2 values were sample-size averaged by grade.

Figure 2.* R^2 *values for Bayesian multilevel logistic regression models predicting scores on various post-promotion rating systems.

2.2.5 Incremental Validity of WAPS Test Scores Over Other WAPS Factors for Predicting Future EPR Ratings

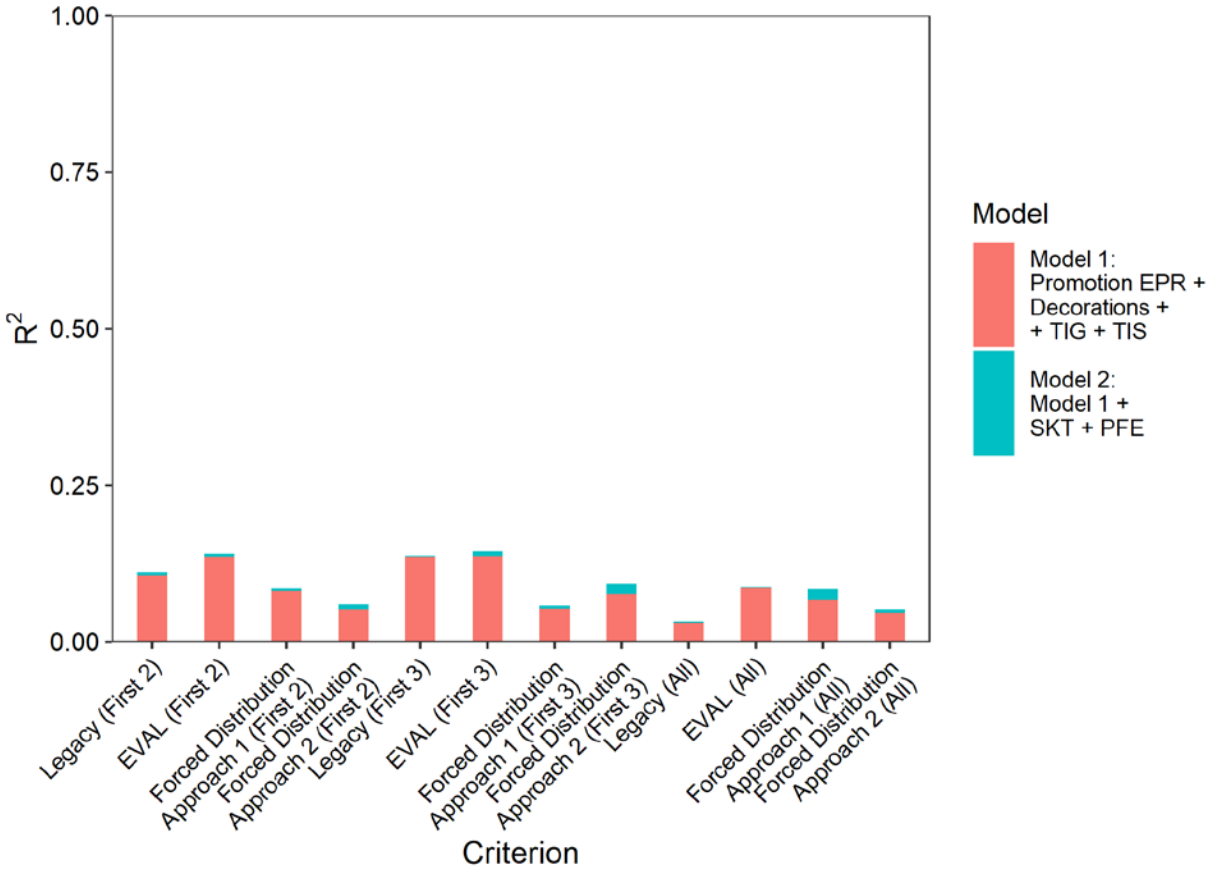
Despite the finding that SKT and PFE scores offered low to negligible amounts of prediction for future EPR ratings, we also planned to examine their incremental validity over other WAPS factors (i.e., control variables): prior EPR ratings (i.e., EPR rating immediately prior to promotion), decorations, TIG, and TIS. Descriptive statistics for these control variables are provided in Appendix A, Exhibit A2. Correlations between these control variables and SKT and PFE scores are given in Exhibit A3 (corrected for multivariate range restriction in Exhibit A4), and between these control variables and the various EPR ratings in Exhibit A5 (corrected for multivariate range restriction in Exhibit A6).

Two Bayesian multilevel logistic regression models were fitted:

- a model including all control variables: prior EPR ratings, decorations, TIG, TIS (Figure 3); and
- a model including only Prior EPR ratings and decorations (estimated due to future phasing out of TIG and TIS from inclusion in WAPS; Figure 4).

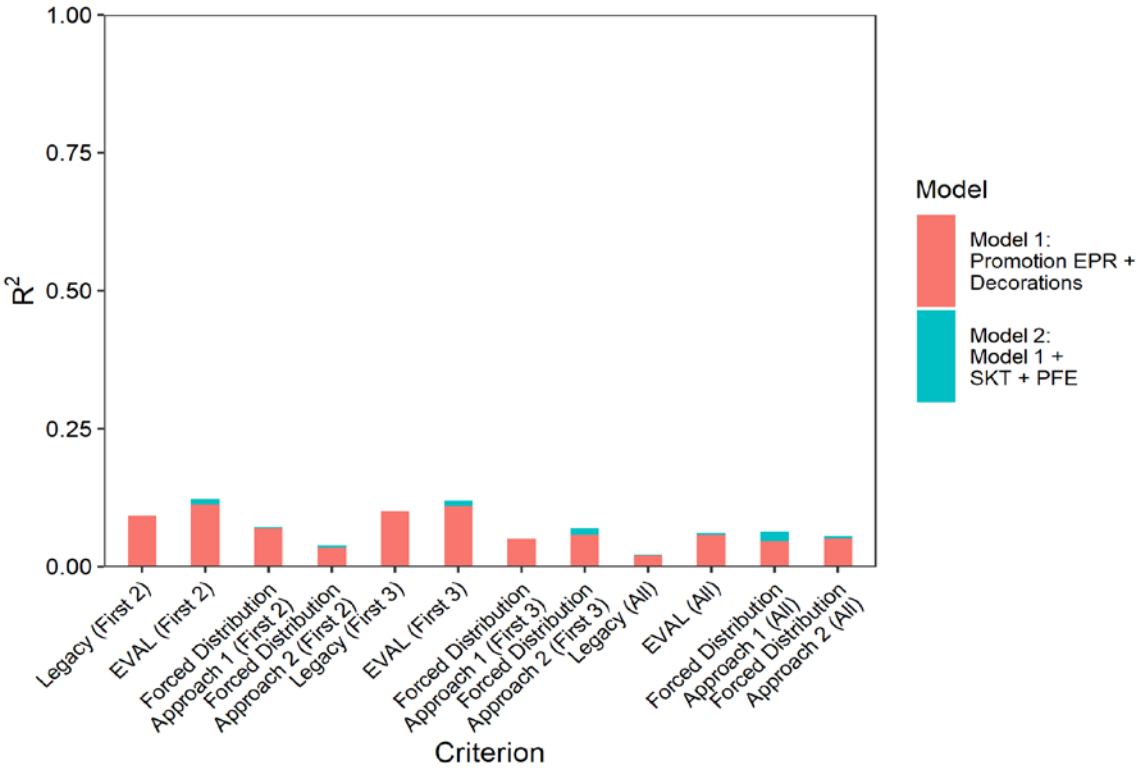
The incremental validity of SKT and PFE scores over the control variables was tested by fitting these two control variable models with SKT and PFE included as additional predictors. Here, the control variables as a group were consistently better predictors of the various EPR ratings than were SKT and PFE scores, having accounted for up to 15% of the variance in post-promotion EPR ratings. Like the SKT and PFE, they were better at predicting the first two or first three post-promotion EPR ratings than all post-promotion EPR ratings. The SKT and PFE scores provided zero or mostly negligible incremental validity over these control variables, regardless of whether this was in reference to all control variables or only prior EPR and decorations.

Additionally, the difference in R^2 estimates between models with all control variables and those with only prior EPR and decorations tends to be around .02. This suggests that among these control variables, the main drivers of predicting future EPR are prior EPR and decorations, not TIG or TIS. This finding would be compatible with the phasing out of TIG and TIS from inclusion in WAPS.



Note. Model 1 includes promotion EPR, decorations, time in grade (TIG), and time in service (TIS) as predictors, and Model 2 includes SKT and PFE as additional predictors. The bar segment for Model 2 represents its increment in R^2 over Model 1. Models were fitted with AFS and cycle as random effects, and predictors as fixed effects and random effects over AFS only. Resulting R^2 values were sample-size averaged by grade.

Figure 3.* R^2 *values for Bayesian multilevel logistic regression models predicting scores on various post-promotion rating systems.



Note. Model 1 includes promotion EPR and decorations as predictors, and Model 2 includes SKT and PFE as additional predictors. The bar segment for Model 2 represents its increment in R^2 over Model 1. Models were fitted with AFS and cycle as random effects, and predictors as fixed effects and random effects over AFS only. Resulting R^2 values were sample-size averaged by grade.

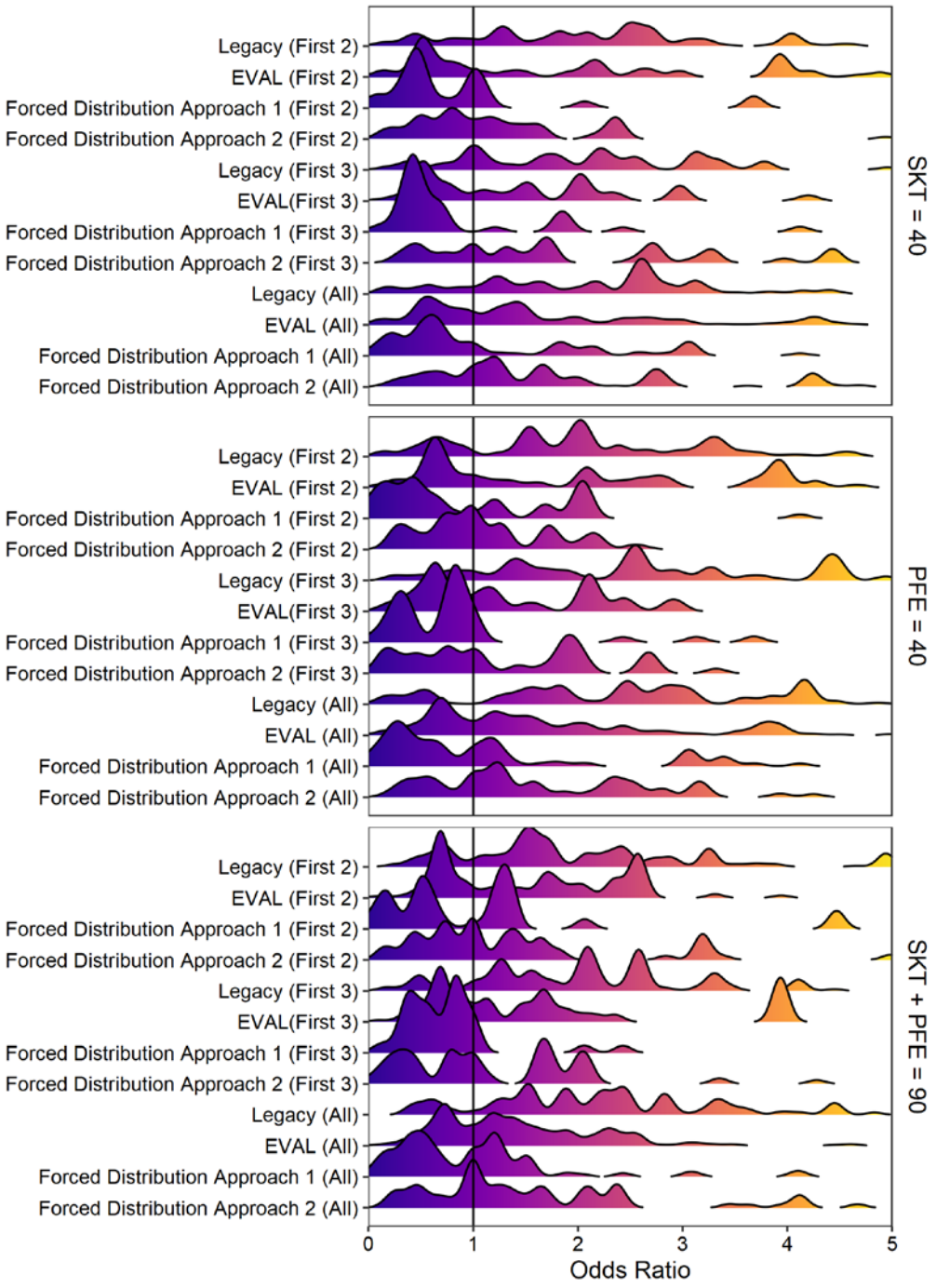
Figure 4.* R^2 *values for Bayesian multilevel logistic regression models predicting scores on various post-promotion rating systems.

2.2.6 Comparison of EPR Ratings among Promoted Airmen Depending on WAPS Cut Scores

Most Airmen in the sample were above the minimum SKT/PFE cut scores of 40 and combined SKT + PFE cut score of 90. To increase sample size for the cut score analysis, outcome variables were predicted in the population data using coefficients obtained from models fitted in the promoted Airmen data by AFS, cycle, and grade. Actual EPR ratings for each Airman were used for this analysis where available, and among those Airmen that did not have an EPR rating (e.g., non-promoted Airmen), their predicted EPR ratings were used instead.

For each cut score (SKT = 40, PFE = 40, SKT + PFE = 90), odds ratios were calculated to compare Airmen who fell above and below the cut score (Figure 5). Odds ratios for passing the cut scores were similar for the individual SKT and PFE cut scores but were slightly lower for the combined SKT and PFE cut score. Nevertheless, similar trends were observed when comparing the pass/fail groups at each cut score. Airmen above the cut score were, on average, over twice as likely to receive the maximum rating on the Legacy EPR system and just under twice as likely to receive the maximum rating on the EVAL EPR system. Odds ratios for some by-AFS, by-cycle,

by-grade combinations were substantially higher than the average, but these are likely to be unreliable due to low sample sizes within such combinations.



Note. The solid vertical line indicates an odds ratio of 1, where Airmen above and below the cut score has the same odds of receiving a maximum rating on the respective rating system.

Figure 5. Distributions of odds ratios across all AFSs, cycles, and grades for receiving a maximum score on various post-promotion ratings systems by cut scores of SKT = 40, PFE = 40, and SKT+PFE = 90.

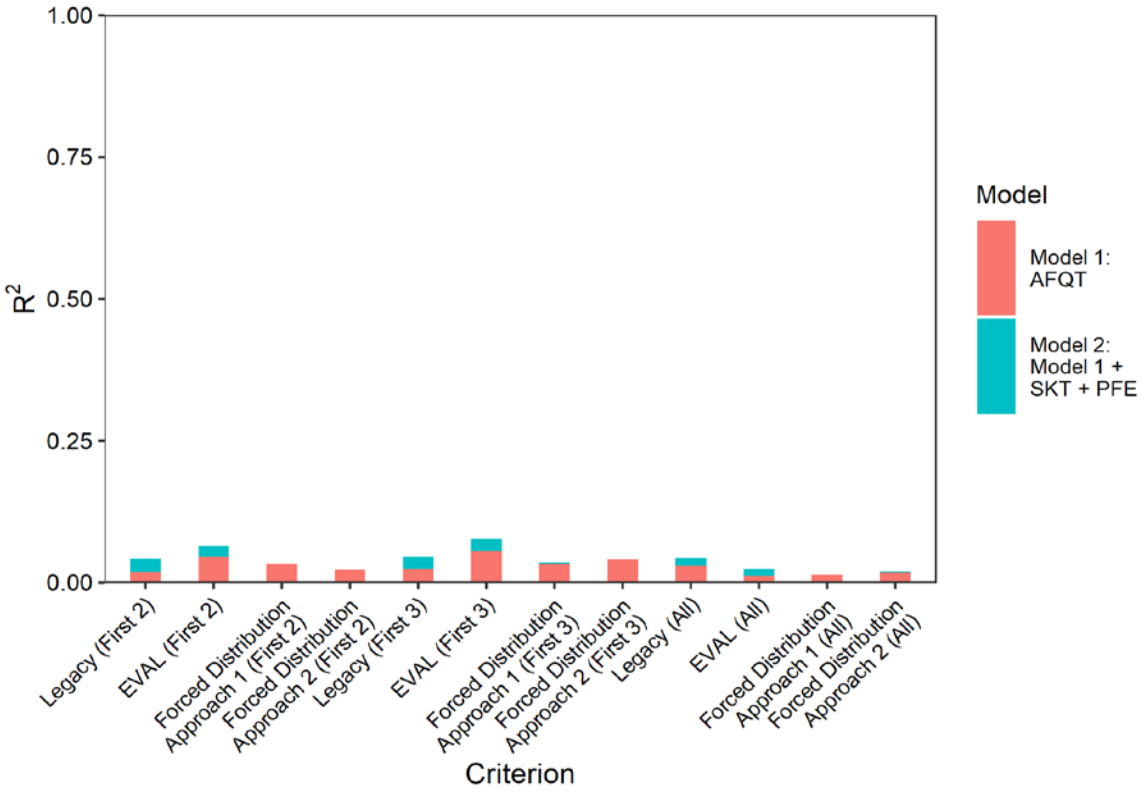
For the forced distribution EPR system, Airmen above the cut scores were about half as likely to receive a rating categorized as maximum when using aggregation approach 1 (Promote Now or Must Promote) and were about 1.5 times as likely to receive a rating categorized as maximum when using aggregation approach 2 (Promote, Promote Now, or Must Promote). This reflects an effect of whether the middle category (Promote) is categorized as maximum in the dichotomizing of the forced distribution EPR ratings.

2.2.7 Validity of WAPS Test Scores for Predicting Future EPR Ratings when Controlled for ASVAB Scores

Besides the control variables described previously, the incremental validity of SKT and PFE scores over ASVAB scores (AFQT and MAGE composites) was tested. Descriptive statistics for the ASVAB scores are provided in Appendix A, Exhibit A7. Correlations between the ASVAB scores and the SKT and PFE scores are shown in Exhibit A8 (corrected for multivariate range restriction in Exhibit A9). Correlations between the ASVAB scores and the SKT and PFE scores are shown in Exhibit A10 (corrected for multivariate range restriction in Exhibit A11).

Regression analyses were conducted separately for the AFQT and for the MAGE composites. For the AFQT, Bayesian multilevel logistic regression models were first fitted with AFQT as a predictor of the various future EPR aggregations, followed by the inclusion of SKT and PFE scores (Figure 6). The AFQT itself provided little to no prediction for future EPR ratings, typically accounting for around 3 to 5% of the variance in aggregated future EPR ratings. Yet, SKT and PFE scores added little to prediction over and above the AFQT, with incremental R^2 of at most .03.

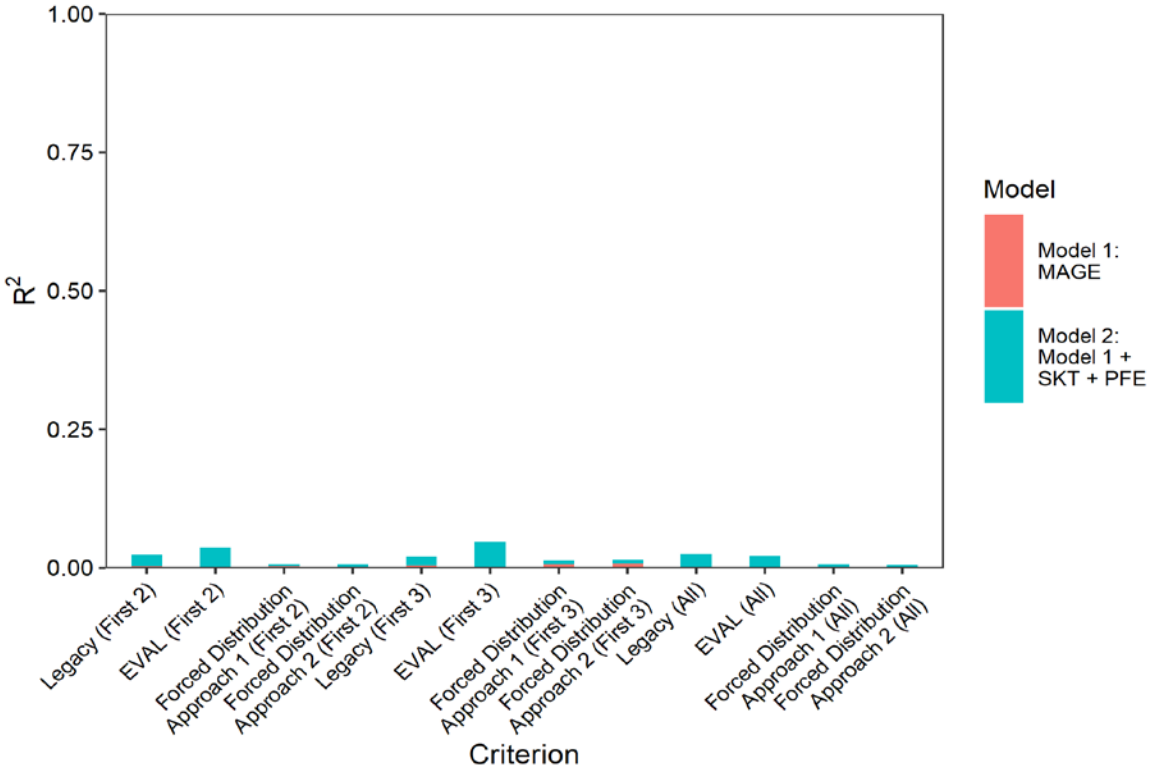
For the MAGE composites, analyses were conducted using single-level logistic regression because the relevant MAGE composites are specific to each focal AFS (with some overlap). In this analysis, separate logistic regression models were fit separated by promoted-to AFS and grade. Cycle was not included in these models due to sample size. For each individual model, each future EPR aggregation was regressed on the AFS-specific MAGE composites. For example, models for Air Traffic Controller (1C1X1) included only General Aptitude (G) as a predictor, whereas models Supply Management (2S0X1) included both General Aptitude (G) and Administrative Aptitude (A) as predictors. These models were then followed up by adding SKT and PFE scores as predictors.



Note. Model 1 includes AFQT as a predictor, and Model 2 includes SKT and PFE as additional predictors. The bar segment for Model 2 represents its increment in *R*2 over Model 1. Models were fitted with AFS and cycle as random effects, and predictors as fixed effects and random effects over AFS only. Resulting *R*2 values were sample-size averaged by grade.

Figure 6.* R^2 *values for Bayesian multilevel logistic regression models predicting scores on various post-promotion rating systems.

Overall, the MAGE composites also offered little to no prediction for future EPR ratings (Figure 7). In most cases (AFS, grade, and criterion combinations), they accounted for no more than 1% of the variance in aggregated future EPR ratings. Although SKT and PFE scores had little incremental validity over the MAGE composites in many cases, there were cases where the SKT and PFE did provide some incremental validity over the MAGE composites—for example when predicting the aggregation of the first three post-promotion forced distribution ratings (aggregation approach 2) for Weather Technician (1W0X1) and grade E-5 group ($\Delta R^2 = .11$). However, we would advise that individual within-AFS and within-grade results be interpreted with caution given the small within-AFS and within-grade sample sizes.



Note. Model 1 includes AFS-specific MAGE line scores as predictors, and Model 2 includes SKT and PFE as additional predictors. The bar segment for Model 2 represents its increment in R^2 over Model 1. Models were fitted by AFS and grade, and resulting R^2 values were sample-size averaged by AFS and grade.

Figure 7.* R^2 *values for single-level logistic regression models predicting scores on various post-promotion rating systems.

2.2.8 Discussion

The results of the criterion-related validity analyses conducted during Task 1 provide little tangible support for the ability of the SKT and PFE to predict Airman performance as operationalized via the Enlistment Performance Reports. Across all models, model fit tended to be relatively poor. Generally, R^2 values ranged mostly from .00 to .05 (multiple correlations from .00 to .22), although in some instances, the R^2 estimate was more substantial, nearing .15 (multiple correlation of .39). In general, models including Promotion EPR, decorations, and/or TIG and TIS tended to demonstrate the highest predictive relations. Typically, the PFE tended to be a marginally stronger predictor than the SKT. That said, these tests offered almost no incremental validity over other predictors (ASVAB, TIS, TIG), and they offered little incremental prediction over the intercept-only baseline models. Despite the relatively low predictive relations observed with the EPR ratings, when including both non-promoted and promoted Airmen and using predicted EPR ratings for non-promoted Airmen, those above the operational cut scores on the SKT and PFE are more likely to have higher EPR ratings. Among promoted Airmen, nearly all of them pass the SKT and PFE cut scores. Still, the primary conclusion to draw is that the SKT and PFE demonstrated weak relations with EPR scores.

The question, then, is why do the WAPS tests not show stronger criterion-related validity? We believe the fault may lie not with the WAPS tests but rather with the criteria they were developed to predict (i.e., the EPR scores). Criterion-related validity is indexed by a correlation – either a bivariate correlation or, in the case of a set of predictors being examined simultaneously, a multiple correlation. Correlations index the covariance (i.e., shared variance) between predictors (X variables) and criteria (Y variables) and, as a consequence, are products of the psychometric characteristics of both the predictors and the criteria.

We believe the evidence is clear that the criteria are the reason the correlations between the WAPS tests and the ratings were low. Two anomalous findings from our analyses are important to highlight. First, none of the ASVAB composites (AFQT, MAGE scores) predicted EPR scores well. This finding runs counter to voluminous evidence supporting the criterion-related validity of ASVAB for predicting job performance in military settings (e.g., Maier & Grafton, 1981; McCloy, Campbell, & Cudeck, 1994; Oppler, McCloy, Peterson, Russell, & Campbell, 2001; Welsh, Kucinkas, & Curran, 1990). Although much of this research focuses on entry-level servicemembers and technical outcomes (e.g., job knowledge, training performance, task performance), researchers have also documented the criterion-related validity of the ASVAB for more advanced ranks against leadership performance, supervisory performance, and other similar outcomes (e.g., Oppler, McCloy, & Campbell, 2001). Given the extensive validity evidence associated with the ASVAB, failure of ASVAB-derived composites to predict job performance is suggestive of problems with the criteria rather than with the predictors.

Second, the EPR ratings had highly restricted variance. For two variables to covary with one another, each variable must vary on its own. At the extreme, a variable that exhibits no variance cannot, by definition, covary with anything. It thus follows that low variance in a measure attenuates its potential covariation with any other measure. Low variability in the legacy ratings led the Air Force to implement the forced distribution ratings. Forced distributions work to limit the types of scores raters can assign to ratees, typically leading to score distributions more palatable for statistical analysis. For the EPR data, however, the percentage constraints at the upper end of the distribution result in the vast majority of Airmen being assigned to the middle category and, ironically, yields less variability in the ratings than was present in the initial legacy ratings the forced distribution ratings were meant to improve (see Tables 9.a-9.c). Low variability plagued all the EPR rating systems, although the EVAL system possessed the most variability.

We caution that the results from the Task 1 analyses do not necessarily mean that the WAPS tests are not useful for predicting performance. It is possible that the SKT and PFE do not predict whatever performance is being evaluated by these EPR systems. For example, we noted early in this report that EPR ratings consider more than just on-the-job performance, capturing off-duty performance, as well. Given the low validity estimates across all predictors (particularly the ASVAB composites), it is quite likely that the EPR ratings do not reflect Airman performance intended to be predicted by the various factors considered during the Airman promotion process. Therefore, we recommend future WAPS validation efforts employ performance rating scales that are developed (a) specifically for the validation effort (i.e., “research-only” scales) and (b) to evaluate components of performance relevant to the predictors being examined. Using such “low stakes” performance ratings typically leads to greater variability and, thus, greater validity than use of performance ratings used for operational personnel decisions (“high stakes”). The psychometric characteristics of the WAPS tests appear strong. We believe validation efforts including research-

only ratings scales (and perhaps other measures of job performance, such as hands-on performance tests, if available) would more properly reflect the validity of these measures.

3.0 TASK 2: ANALYZE ARCHIVAL DATA BASED ON EXAMINEES' FIRST-TIME AND REPEATED ITEM EXPOSURE

In Task 2, we conducted a series of analyses to investigate the effects of item exposure on SKT and PFE performance, and the moderating effect of item exposure on the associations between SKT and PFE performance and demographic and experience factors. We organized Task 2 into five broad analyses:

- Analysis 1: Item use history as a moderator of relations between demographic/experience factors and SKT and PFE performance
- Analysis 2: Individuals' previous exposure to test items as a moderator of relations between demographic/experience factors and SKT and PFE performance
- Analysis 3: Repeat candidate status as a moderator of item-level statistics
- Analysis 4: Differential item functioning by first-time and repeat examinees
- Analysis 5: Moderating effect of tenure on mean SKT and PFE score differences by sex, race, and ethnicity.

3.1 Analysis 1: Item Use History

In Analysis 1, we analyzed results comparing demographic differences in WAPS exam performance (SKT and PFE) based on first time and repeated item use. Demographic variables included sex, race, and ethnicity. We also examined the moderating effect of item use history on the magnitude of the relations between SKT/PFE scores and three variables: Armed Forces Qualification Test (AFQT) scores, time in service (TIS), and time in grade (TIG).

3.1.1 Method

Sample. The sample used in this study consisted of Airmen who (a) were seeking promotion to grade E-5, E-6, or E-7, (b) took a relevant promotion test (a PFE and/or an SKT associated with one of the following Air Force Specialty Codes (AFSC): 1C1X1, 1P0X1, 1W0X1, 2S0X1, 2W1X1, 3E7X1, 3P0X1, or 4N0X1) for the grade above their current grade, and (c) were tested in the 2011 through 2016 promotion cycles. For data-quality assurance, we excluded examinees with incomplete item response data. We merged Airmen's item-level data with their prior testing records to identify those who had previously taken the exam for which they were sitting. We also merged Airmen's item responses with their AFQT scores. In the case that multiple AFQT scores were on record for a given Airman, we used the average of their unique scores as their AFQT score in our analyses. We used items' usage histories to identify which items in each test administration were new and which had been used in at least one previous administration.

Tables 10.a and 10.b present the maximum sample sizes by group for Analysis 1. Analyses 2-5 were also based on this sampling frame, although the specifics of analyses 2-5 dictate that the sample sizes will not be the same in each analysis. (We describe the unique inclusion criteria in the method section for each analysis.)

Table 10.a. Sampling Frame for Analyses 1-5: SKT

SKT	Revision	Men	Women	White	Black	Asian	Am. Ind.	Hispanic	Non-Hispanic	AFQT *N*	TIS *N*	TIG *N*
1C151	17	333	67	310	46	14	0	5	442	396	397	397
1C151	62	504	99	475	69	16	4	11	616	601	602	602
1C151	63	570	101	518	87	16	2	12	693	665	666	666
1C151	64	597	99	553	77	17	4	14	722	690	692	692
1C151	65	656	106	602	81	22	4	15	808	746	749	749
1C151	66	673	125	612	103	21	4	12	818	777	779	779
1C171	17	394	132	396	76	8	0	39	551	509	517	517
1C171	63	438	104	386	80	17	6	39	525	525	531	531
1C171	64	525	108	444	92	19	6	41	616	612	621	621
1C171	65	478	106	435	68	17	4	33	589	568	574	574
1C171	66	437	97	407	69	14	2	23	540	516	521	521
1P051	2	226	69	203	66	6	3	11	350	281	288	288
1P051	62	244	66	205	71	10	5	4	363	304	307	307
1P051	63	293	56	237	74	14	2	11	385	341	346	346
1P051	64	295	56	244	67	11	2	11	401	344	348	348
1P051	65	335	58	281	68	12	0	10	451	382	387	387
1P051	66	334	44	267	61	19	3	9	406	369	370	370
1P071	2	403	106	309	120	8	11	47	530	479	502	502
1P071	63	392	112	320	121	11	8	36	511	478	503	503
1P071	64	401	117	329	124	9	4	34	530	491	515	515
1P071	65	396	113	316	118	8	3	37	524	470	502	502
1P071	66	376	101	299	113	9	4	25	483	448	474	474
1W051	16	138	62	167	20	5	2	3	238	197	197	197
1W051	62	193	79	224	24	10	2	4	311	272	272	272
1W051	63	260	97	296	35	11	2	3	397	357	357	357
1W051	64	273	94	297	41	10	2	7	420	362	364	364
1W051	65	366	100	380	47	13	0	7	500	455	458	458
1W051	66	357	94	366	50	9	0	12	468	431	433	433
1W071	16	359	95	365	35	9	2	25	510	447	451	451
1W071	63	405	96	418	35	13	2	25	524	485	493	493
1W071	64	427	101	434	43	10	2	26	549	521	526	526

Table 10.a. (Continued)

SKT	Revision	Men	Women	White	Black	Asian	Am. Ind.	Hispanic	Non-Hispanic	AFQT *N*	TIS *N*	TIG *N*
1W071	65	417	101	422	42	9	4	26	535	511	516	516
1W071	66	435	93	423	51	10	5	22	541	513	521	521
2S051	16	559	371	416	377	47	15	44	1064	919	922	922
2S051	62	603	380	441	395	54	17	47	1120	973	981	981
2S051	63	596	323	446	336	51	16	46	1029	889	902	902
2S051	64	608	318	479	303	56	24	41	1013	909	924	924
2S051	65	753	340	580	342	66	19	51	1186	1053	1069	1069
2S051	66	679	326	506	327	62	14	40	1089	966	975	975
2S071	16	940	708	636	697	76	10	196	1810	1589	1640	1640
2S071	63	936	720	624	709	100	12	157	1716	1596	1652	1652
2S071	64	928	739	618	709	106	12	154	1700	1589	1644	1644
2S071	65	871	648	563	652	98	12	133	1540	1427	1506	1506
2S071	66	841	569	543	593	90	10	124	1395	1341	1401	1401
2W151	17	680	105	551	120	35	10	29	915	777	783	783
2W151	62	805	130	645	174	35	5	31	1016	923	931	931
2W151	63	789	124	651	157	37	7	37	991	894	906	906
2W151	64	791	124	654	158	31	6	35	983	899	912	912
2W151	65	938	139	783	173	33	9	38	1167	1030	1039	1039
2W151	66	960	146	779	177	43	14	37	1156	1046	1050	1050
2W171	17	1439	102	1060	271	45	9	148	1575	1496	1535	1535
2W171	63	1405	120	1003	284	58	10	136	1516	1469	1516	1516
2W171	64	1481	137	1091	283	57	12	147	1609	1560	1605	1605
2W171	65	1505	151	1086	303	72	19	135	1663	1589	1642	1642
2W171	66	1370	132	994	267	72	14	111	1505	1431	1479	1479
3P051	11	2719	685	2325	746	72	43	125	4169	3330	3370	3370
3P051	62	3435	842	2880	972	101	47	154	4783	4198	4247	4247
3P051	63	3595	835	3035	964	105	50	150	5062	4351	4401	4401
3P051	64	3888	826	3325	931	96	55	135	5180	4622	4666	4666
3P051	65	4324	885	3688	993	112	68	136	5662	5029	5066	5066
3P051	66	4270	838	3606	976	127	60	122	5566	4825	4845	4845
3P071	11	3299	471	2647	666	84	36	311	4088	3632	3748	3748

Table 10.a. (Continued)

SKT	Revision	Men	Women	White	Black	Asian	Am. Ind.	Hispanic	Non-Hispanic	AFQT *N*	TIS *N*	TIG *N*
P071	63	3875	599	3163	817	104	43	290	4564	4297	4445	4445
3P071	64	3982	661	3241	871	107	45	289	4786	4411	4556	4556
3P071	65	3965	649	3196	898	98	43	276	4683	4343	4527	4527
3P071	66	3823	599	3061	856	105	47	272	4506	4162	4345	4345
4N051	17	254	335	361	129	40	10	21	672	582	584	584
4N051	62	320	443	479	161	45	13	25	845	754	759	759
4N051	63	377	554	595	187	67	11	22	1022	920	927	927
4N051	64	470	556	670	210	60	6	26	1102	1010	1017	1017
4N051	65	531	549	714	205	68	8	29	1144	1047	1050	1050
4N051	66	579	562	764	207	62	12	34	1165	1097	1098	1098
4N071	17	480	616	611	273	58	8	122	1155	1069	1092	1092
4N071	63	551	679	707	297	62	8	114	1203	1197	1226	1226
4N071	64	570	644	693	292	69	6	120	1182	1176	1202	1202
4N071	65	622	684	749	308	81	8	112	1256	1265	1295	1295
4N071	66	594	657	707	286	86	13	98	1209	1196	1228	1228
3E751	17	485	11	392	63	11	3	17	565	490	493	493
3E751	62	506	8	422	57	8	0	12	556	504	505	505
3E751	63	478	5	386	62	4	4	7	533	474	477	477
3E751	64	497	9	391	77	5	7	8	569	505	505	505
3E751	65	565	9	445	82	7	5	9	689	554	555	555
3E751	66	617	16	469	95	13	8	13	691	610	612	612
3E771	17	537	11	398	83	12	6	48	582	540	547	547
3E771	63	605	9	469	73	6	6	47	600	595	610	610
3E771	64	659	9	506	87	10	5	48	676	638	655	655
3E771	65	613	12	468	78	13	8	34	667	607	623	623
3E771	66	611	11	465	84	13	6	34	634	600	615	615

Table 10.b. Sampling Frame for Analyses 1-5: PFE

PFE	Revision	Men	Women	White	Black	Asian	Am. Ind.	Hispanic	Non-Hispanic	AFQT *N*	TIS *N*	TIG *N*
00035A	58	12411	2426	11110	2169	465	166	484	17427	14367	14451	14451
00035B	58	8970	3451	8190	2701	516	140	467	14314	11985	12098	12098
00035A	59	14236	2499	12427	2573	558	161	510	18485	16244	16356	16356
00035B	59	12490	4594	11571	3558	665	192	592	16492	16575	16704	16704
00035A	60	12920	2323	11350	2357	523	138	427	16226	14794	14926	14926
00035B	60	15127	4481	13698	3686	772	210	583	20723	18986	19152	19152
00035	61	30317	7120	26873	6334	1347	392	1018	41105	36468	36739	36739
00035	62	33353	7468	29365	6674	1501	398	1047	44800	39006	39260	39260
00035	63	33651	7605	29125	6988	1606	402	1047	43861	38851	39064	39064
00036A	43	12334	3832	10693	3167	575	167	1361	17069	15280	15698	15698
00036B	43	15300	3798	13126	3163	737	171	1699	19757	18039	18559	18559
00036A	60	13126	3595	11268	3183	663	177	1101	16515	15944	16372	16372
00036B	60	17655	4193	15455	3393	858	222	1413	21423	20694	21236	21236
00036	61	31668	7803	27592	6663	1545	377	2341	39667	37440	38344	38344
00036	62	29546	7292	25874	6156	1412	376	1952	37821	34848	35863	35863
00036	63	28101	6609	24246	5912	1396	353	1711	34644	32654	33569	33569
00037A	43	7077	2083	6098	1710	200	72	761	10103	8798	8957	8957
00037B	43	8822	1927	7656	1708	229	65	880	11601	10421	10581	10581
00037A	60	7029	2060	5697	1869	276	75	876	8850	8654	8944	8944
00037B	60	9569	2187	8057	1904	324	83	1086	11290	11291	11584	11584
00037	61	18397	4622	15154	4106	766	181	2219	22600	21933	22673	22673
00037	62	19112	4895	15697	4355	885	200	2311	22813	22736	23619	23619
00037	63	17757	4194	14370	4008	852	206	2052	21174	20688	21504	21504

Analyses. Analysis 1 primarily concerned item use history (i.e., whether the item was new or reused, and for reused items, the duration since last use). Within an exam administration year (promotion board cycle), each item on each revision of an SKT or PFE was coded according to its use history:

- New item (First-time)
- Repeat item that had been used in the previous year (Repeat item, consecutive cycles)
- Repeat item that had been used 2 years ago but not in the previous year (Repeat item, rested one year)
- Repeat item that had been used 3 or more years ago but not in the previous 2 years (Repeat item, rested 2 or more years)

We included items in analyses if we were able to definitively code the item history. For example, if an item was administered in 2011 and again in 2012, its 2012 status was coded as a repeat item, consecutive cycles. However, if an item's initial status code was a value other than "new" or "experimental" and the prior use could not be determined from available data, then the item use history was set to missing. For this reason, the number of items in each condition differs across years.

We merged candidates' SKT and PFE item responses for each year (revision) with promotion board cycle information (sex, race, ethnicity, AFQT scores, TIS scores, and TIG scores). We combined this information with the item within administration-use statuses.

For each candidate within each SKT or PFE administration (revision), we computed up to four scores based on the item use histories indicated above. Because the number of items in each condition varies widely across years and test forms, scores for a given condition were computed only if there were more than five items on a test form in that condition.

The variation in the number of items in each item use history can influence the reliability of a score calculated from those items (e.g., some scores might be based on 40 items while others might be based on 10 or fewer items). We calculated the internal consistency reliability of the scores from each item use history condition in each revision of each SKT or PFE to account for differences in reliability.

Within each revision of each SKT and PFE, we calculated mean scores for each item use condition for men and women as well as for the following demographic groups: White, Black, Asian, American Indian or Alaska Native (AI/AN), Hispanic, and Non-Hispanic. We calculated standardized mean difference effect sizes (*d* values) in exam scores based on sex, race, and ethnic groups for any condition with more than 20 individuals in each group. We meta-analyzed the *d* values using the *psychmeta* package for R. We corrected for unreliability in scores to account for differences in the numbers of items in each condition, and we explored the potential moderating effect of item use condition on differences by sex, race, and ethnicity.

Finally, we computed correlations between scores in each item use condition and AFQT scores, TIS scores, and TIG scores. We meta-analyzed those correlations using the *psychmeta* package for R (correcting for unreliability in scores) and explored the potential moderating effect of item use condition.

3.1.2 Results

Sex Differences by Item Status. Table 11 displays mean differences in scores on SKT items[6] as a function of item use history and sex. The mean effects tended to be small to moderate, with an overall sex difference (*d*) of 0.19. The overall effect sizes by SKT ranged from -0.07 (scores for 2S071 women were slightly higher, on average, than scores for 2S071 men) to 0.39 (scores for 2W151 men were moderately higher, on average, than scores for 2W151 women).
Focusing on the results by item condition, the magnitude of the group differences by sex were not moderated by item use history. The *d* values are very similar across conditions, and the variance in effect sizes within levels of item use history is comparable to the variance in effect sizes overall. For first time use items, the mean sex difference was 0.19 (corrected for unreliability $\delta = 0.27$, SD $\delta = 0.21$)[7], which is not much different from the magnitude of the sex difference for other item use histories. For example, the difference for repeated items that sat out one cycle since their previous use was 0.22 ($\delta = 0.23$, SD $\delta = 0.17$).

Table 12 displays average differences in scores on PFE items as a function of item use history and sex. The average effects were near zero, with an overall sex difference (*d*) of 0.003. The overall effect sizes by PFE (00035, 00036, and 00037) ranged from -0.02 (00037) to 0.02 (00035).

Focusing on the results by item condition, the magnitude of the group differences by sex was not moderated by item use history. The *d* values are near zero in each item use condition, and the variance in effect sizes within levels of item use history is comparable to the variance in effect sizes overall. For first time use items, the mean group difference was 0.02 ($\delta = 0.04$, SD $\delta = 0.17$), which is not much different from the magnitude of the group difference for other item use histories. For example, the mean for repeated items that were rested for two or more cycles was -0.02 ($\delta = -0.02$, SD $\delta = 0.19$).

Race Differences by Item Status. Tables 13-15 present results for score differences on SKT items by race (White-Black, White-Asian, and White-AI/AN). The largest race differences were associated with White-Black comparisons (Table 13), with an overall White-Black *d* value of 0.25 ($\delta = 0.34$). The overall *d* comparing average scores for White Airmen and average scores for Asian Airmen was 0.11 ($\delta = 0.13$), and the White-AI/AN *d* was 0.15 ($\delta = 0.25$).

[6] Throughout this section we refer to "scores on SKT items" as opposed to SKT scores to maintain the distinction that the scores we computed for these analyses are based on subsets of items from the SKT. This distinction is also used for scores on PFE items.

[7] δ = sample-size-weighted standardized mean difference corrected for unreliability in item sets; sd δ = standard deviation of corrected effect sizes not attributable to sampling error;

Table 11. Sex Differences on SKT Items by Item Use History and SKT

Item Condition		SKT	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	mean_delta	sd_delta	CI_LL_95	CI_UL_95
All	All	188	247503	0.19	0.0355	0.0051	0.0305	0.23	0.20	0.20	0.26	
Repeat, consecutive cycles	All	32	36024	0.18	0.0467	0.0056	0.0411	0.17	0.21	0.09	0.25	
Repeat, sat out one cycle	All	56	79715	0.22	0.0347	0.0048	0.0300	0.23	0.17	0.18	0.28	
Repeat, sat out two+ cycles	All	34	45804	0.16	0.0425	0.0050	0.0375	0.25	0.28	0.15	0.36	
First time	All	66	85960	0.19	0.0283	0.0052	0.0232	0.27	0.21	0.21	0.32	
All	1C151	15	10571	0.07	0.0046	0.0113	-0.0067	0.10	0.00	0.05	0.15	
All	1C171	13	7496	0.05	0.0047	0.0116	-0.0069	0.05	0.00	0.00	0.11	
All	1P051	14	5089	0.07	0.0239	0.0221	0.0017	0.10	0.08	-0.03	0.23	
All	1P071	14	7015	0.08	0.0174	0.0117	0.0057	0.13	0.09	0.03	0.24	
All	1W051	14	5504	0.30	0.0191	0.0142	0.0049	0.45	0.05	0.35	0.56	
All	1W071	13	6706	0.29	0.0172	0.0128	0.0044	0.40	0.06	0.30	0.50	
All	2S051	15	14754	-0.02	0.0068	0.0046	0.0022	-0.04	0.06	-0.11	0.02	
All	2S071	14	21825	-0.07	0.0032	0.0026	0.0006	-0.11	0.00	-0.15	-0.07	
All	2W151	14	13946	0.39	0.0099	0.0088	0.0011	0.53	0.09	0.44	0.62	
All	2W171	12	18803	0.32	0.0381	0.0083	0.0298	0.45	0.20	0.30	0.60	
All	3P051	12	56526	0.26	0.0107	0.0015	0.0093	0.39	0.11	0.31	0.47	
All	3P071	11	49112	0.31	0.0341	0.0019	0.0321	0.44	0.24	0.27	0.61	
All	4N051	14	13981	0.11	0.0115	0.0041	0.0074	0.19	0.10	0.11	0.28	
All	4N071	13	16175	0.05	0.0057	0.0032	0.0024	0.07	0.09	-0.01	0.15	

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean delta = sample-size-weighted standardized mean difference corrected for unreliability in item sets; sd delta = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 12. Sex Differences on PFE Items by Item Use History and PFE

Item Condition	PFE	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	mean_delta	sd_delta	CI_LL_95	CI_UL_95
All	All	40	1091961	0.00	0.0124	0.0002	0.0122	0.01	0.16	-0.04	0.07
Repeat, sat out one cycle	All	9	299510	0.00	0.0096	0.0002	0.0094	0.00	0.14	-0.11	0.11
Repeat, sat out two+ cycles	All	9	299510	-0.02	0.0163	0.0002	0.0161	-0.02	0.19	-0.16	0.13
First time	All	22	492941	0.02	0.0129	0.0003	0.0126	0.04	0.17	-0.04	0.11
All	00035	14	437386	0.02	0.0046	0.0002	0.0044	0.03	0.10	-0.03	0.09
All	00036	13	406890	0.00	0.0237	0.0002	0.0235	0.02	0.23	-0.12	0.16
All	00037	13	247685	-0.02	0.0094	0.0003	0.0091	-0.03	0.14	-0.11	0.06

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean delta = sample-size-weighted standardized mean difference corrected for unreliability in item sets; sd delta = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 13. White-Black Differences on SKT Items by Item Use History and SKT

Item Condition	SKT	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	mean_delta	sd_delta	CI_LL_95	CI_UL_95
All	All	216	229814	0.25	0.0223	0.0064	0.0159	0.34	0.16	0.31	0.36
Repeat, consecutive cycles	All	38	34185	0.28	0.0237	0.0077	0.0160	0.32	0.15	0.26	0.38
Repeat, sat out one cycle	All	64	73655	0.29	0.0223	0.0059	0.0164	0.34	0.14	0.30	0.38
Repeat, sat out two+ cycles	All	37	41404	0.18	0.0207	0.0060	0.0148	0.31	0.22	0.22	0.39
First time	All	77	80570	0.23	0.0186	0.0065	0.0120	0.37	0.16	0.32	0.41
All	1C151	15	9507	0.12	0.0083	0.0142	-0.0059	0.16	0.00	0.08	0.24
All	1C171	13	6485	0.18	0.0145	0.0154	-0.0009	0.25	0.00	0.16	0.34
All	1P051	14	4497	0.28	0.0177	0.0192	-0.0015	0.44	0.00	0.34	0.54
All	1P071	14	6054	0.18	0.0159	0.0117	0.0041	0.28	0.07	0.18	0.38
All	1W051	14	5067	0.35	0.0191	0.0280	-0.0089	0.52	0.00	0.43	0.61
All	1W071	13	6025	0.35	0.0295	0.0260	0.0035	0.49	0.16	0.32	0.65
All	2S051	15	12419	0.09	0.0095	0.0051	0.0045	0.14	0.08	0.06	0.21
All	2S071	14	17496	0.04	0.0068	0.0032	0.0036	0.05	0.08	-0.01	0.12
All	2W151	14	12234	0.28	0.0091	0.0075	0.0016	0.38	0.08	0.30	0.47
All	2W171	12	15864	0.28	0.0196	0.0045	0.0150	0.39	0.12	0.29	0.49
All	3E751	14	7034	0.38	0.0433	0.0156	0.0277	0.57	0.11	0.44	0.69
All	3E771	14	7768	0.33	0.0272	0.0146	0.0126	0.47	0.15	0.34	0.61
All	3P051	12	50896	0.32	0.0136	0.0014	0.0122	0.49	0.12	0.41	0.57
All	3P071	11	43530	0.27	0.0139	0.0015	0.0124	0.37	0.16	0.26	0.49
All	4N051	14	11884	0.25	0.0205	0.0067	0.0138	0.41	0.14	0.30	0.52
All	4N071	13	13054	0.19	0.0053	0.0048	0.0004	0.31	0.05	0.23	0.38

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean delta = sample-size-weighted standardized mean difference corrected for unreliability in item sets; sd delta = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 14. White-Asian Differences on SKT Items by Item Use History and SKT

Item Condition	SKT	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	mean_delta	sd_delta	CI_LL_95	CI_UL_95
All	All	112	142275	0.11	0.0230	0.0149	0.0082	0.13	0.15	0.08	0.17
Repeat, consecutive cycles	All	19	19350	0.10	0.0200	0.0163	0.0037	0.10	0.12	0.01	0.20
Repeat, sat out one cycle	All	34	47012	0.13	0.0225	0.0141	0.0084	0.13	0.14	0.07	0.20
Repeat, sat out two+ cycles	All	21	26431	0.07	0.0266	0.0154	0.0113	0.04	0.22	-0.10	0.19
First time	All	38	49482	0.11	0.0233	0.0148	0.0085	0.17	0.15	0.08	0.25
All	1C151	7	4395	-0.09	0.0169	0.0482	-0.0313	-0.13	0.00	-0.30	0.03
All	2S051	15	8302	-0.09	0.0193	0.0193	0.0000	-0.15	0.00	-0.25	-0.04
All	2S071	14	9571	-0.09	0.0240	0.0121	0.0118	-0.13	0.16	-0.25	0.00
All	2W151	14	10433	0.01	0.0165	0.0293	-0.0128	0.01	0.00	-0.09	0.12
All	2W171	12	13261	0.14	0.0292	0.0168	0.0124	0.18	0.14	0.04	0.32
All	3P051	12	40787	0.17	0.0126	0.0099	0.0028	0.25	0.04	0.16	0.35
All	3P071	11	35341	0.17	0.0071	0.0102	-0.0031	0.24	0.00	0.16	0.31
All	4N051	14	9993	0.13	0.0109	0.0181	-0.0071	0.21	0.00	0.11	0.30
All	4N071	13	10192	0.16	0.0209	0.0150	0.0058	0.28	0.08	0.15	0.41

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean delta = sample-size-weighted standardized mean difference corrected for unreliability in item sets; sd delta = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 15. White-AI/AN Differences on SKT Items by Item Use History and SKT

Item Condition	SKT	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	mean_delta	sd_delta	CI_LL_95	CI_UL_95
All	All	27	76925	0.15	0.0357	0.0211	0.0146	0.25	0.11	0.15	0.34
Repeat, consecutive cycles	All	3	7277	0.28	0.0871	0.0208	0.0663	0.32	0.33	-0.62	1.26
Repeat, sat out one cycle	All	9	27229	0.25	0.0215	0.0205	0.0010	0.30	0.06	0.16	0.44
Repeat, sat out two+ cycles	All	5	13986	0.03	0.0255	0.0203	0.0052	0.07	0.07	-0.29	0.43
First time	All	10	28433	0.08	0.0258	0.0220	0.0038	0.18	0.00	0.00	0.35
All	2S051	4	2012	0.03	0.0062	0.0439	-0.0377	0.04	0.00	-0.16	0.24
All	3P051	12	40205	0.22	0.0425	0.0181	0.0244	0.37	0.13	0.22	0.52
All	3P071	11	34708	0.07	0.0195	0.0232	-0.0037	0.13	0.00	0.01	0.24

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean delta = sample-size-weighted standardized mean difference corrected for unreliability in item sets; sd delta = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Item status did not moderate the magnitude of group differences. Across item history conditions, the White-Black and White-Asian *d* values were similar in magnitude, and there was no noticeable reduction in variance (SD δ) within item history condition. In the White-AI/AN comparisons (Table 15), there were differences in effect sizes across item history conditions, with smaller group differences associated with new items and items that had rested for 2 or more years and larger group differences associated with items used in consecutive years or after resting only one year. However, the numbers of conditions and total sample sizes for the White-AI/AN comparisons were much smaller than other analyses, and the confidence intervals overlap substantially. As such, item use history does not seem to moderate the White-AI/AN group difference.

Tables 16-18 present group difference results for the PFE. As with the SKT results, the group differences were generally small and not moderated by item use history.

Ethnicity Differences by Item Status. Tables 19 and 20 present group difference results on SKT and PFE items by ethnicity (Hispanic or non-Hispanic). The differences were very small, with *d* values near zero. Results corrected for unreliability were slightly larger, but the δ values were still below 0.10. The magnitude of the *d* and δ values were similar across item use conditions, and the sd δ values were not reduced compared with the SD δ values overall.

AFQT-Performance Correlations by Item Status. Tables 21 and 22 display the meta-analysis results for the AFQT-score correlations for SKT and PFE items, respectively. Scores across SKT and PFE items were correlated with AFQT scores, with the weighted average AFQT-SKT correlation of .19 (ρ = .28) and the weighted average AFQT-PFE correlation of .20 (ρ = .31). Item use history did not act as a substantive moderator, with nearly equal ρ values across conditions and SD ρ values that are as large within item use condition as they are overall.

Tenure-Performance Correlations by Item Status. Tables 23-24 present results of correlations between tenure (time-in-service, time-in-grade) and scores on SKT and PFE items. The correlation between performance on SKT items and TIS was very small, with an overall weighted average correlation of .01 (ρ = .02). The correlation between performance on PFE items and time-in-service was small and negative, with an overall weighted average correlation of -.07 (ρ = -.10). Within item history conditions, the correlations between scores and TIS were consistent.

The pattern of results was similar for correlations with TIG. The overall correlations between TIG and performance on SKT and PFE items were near zero, and the magnitude of the correlations did not change as a function of item use history.

3.1.3 Discussion

The results from Analysis 1 are consistent in suggesting that item use history does not moderate relations between item performance and demographic and experience factors. The magnitude of *d* values by sex, race, and ethnicity was consistent across item use history conditions.

A few observations stood out in Analysis 1. First, there were some differences across AFSs in the magnitude of group differences, particularly by sex. For example, there was very little difference in the average performance of men and women within 1C1X1, 1P0X1, 2S0X1, and 4N0X1. The *d* values by sex were more noticeable in 1W0X1, 2W1X1, and 3P0X1.

Table 16. White-Black Differences on PFE Items by Item Use History and PFE

Item Condition	PFE	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	mean_delta	sd_delta	CI_LL_95	CI_UL_95
All	All	40	946884	0.12	0.0073	0.0003	0.0070	0.18	0.13	0.14	0.22
Repeat, sat out one cycle	All	9	259492	0.12	0.0058	0.0002	0.0056	0.17	0.10	0.09	0.25
Repeat, sat out two+ cycles	All	9	259492	0.11	0.0062	0.0002	0.0059	0.17	0.11	0.09	0.25
First time	All	22	427900	0.13	0.0095	0.0003	0.0092	0.18	0.16	0.11	0.25
All	00035	14	386338	0.13	0.0041	0.0002	0.0039	0.20	0.10	0.14	0.26
All	00036	13	352777	0.14	0.0080	0.0002	0.0078	0.20	0.14	0.12	0.29
All	00037	13	207769	0.07	0.0096	0.0004	0.0092	0.10	0.14	0.01	0.19

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean delta = sample-size-weighted standardized mean difference corrected for unreliability in item sets; sd delta = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 17. White-Asian Differences on PFE Items by Item Use History and PFE

Item Condition	PFE	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	mean_delta	sd_delta	CI_LL_95	CI_UL_95
All	All	40	800339	0.06	0.0081	0.0011	0.0070	0.10	0.12	0.05	0.14
Repeat, sat out one cycle	All	9	219606	0.10	0.0111	0.0008	0.0103	0.15	0.15	0.03	0.27
Repeat, sat out two+ cycles	All	9	219606	0.08	0.0021	0.0008	0.0013	0.12	0.04	0.08	0.17
First time	All	22	361127	0.03	0.0085	0.0014	0.0071	0.04	0.13	-0.02	0.11
All	00035	14	329060	0.03	0.0071	0.0009	0.0061	0.05	0.12	-0.02	0.12
All	00036	13	299570	0.10	0.0081	0.0009	0.0072	0.15	0.12	0.07	0.23
All	00037	13	171709	0.06	0.0079	0.0017	0.0061	0.10	0.12	0.02	0.18

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean delta = sample-size-weighted standardized mean difference corrected for unreliability in item sets; sd delta = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 18. White-AI/AN Differences on PFE Items by Item Use History and PFE

Item Condition	PFE	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	mean_delta	sd_delta	CI_LL_95	CI_UL_95
All	All	40	770215	0.11	0.0082	0.0040	0.0042	0.17	0.09	0.13	0.21
Repeat, sat out one cycle	All	9	211181	0.12	0.0060	0.0032	0.0028	0.18	0.07	0.09	0.26
Repeat, sat out two+ cycles	All	9	211181	0.12	0.0110	0.0032	0.0078	0.18	0.13	0.06	0.30
First time	All	22	347853	0.11	0.0088	0.0049	0.0039	0.16	0.09	0.10	0.23
All	00035	14	317255	0.14	0.0040	0.0032	0.0008	0.21	0.05	0.15	0.26
All	00036	13	287733	0.08	0.0073	0.0032	0.0041	0.12	0.08	0.05	0.20
All	00037	13	165227	0.12	0.0166	0.0066	0.0100	0.19	0.15	0.08	0.31

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean delta = sample-size-weighted standardized mean difference corrected for unreliability in item sets; sd delta = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 19. Non-Hispanic-Hispanic Differences on SKT Items by Item Use History and SKT

Item Condition	SKT	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	mean_rho	sd_rho	CI_LL_95	CI_UL_95
All	All	159	259804	0.03	0.0232	0.0143	0.0089	0.03	0.13	0.00	0.07
Repeat, consecutive cycles	All	28	37126	0.00	0.0245	0.0179	0.0066	-0.01	0.11	-0.08	0.07
Repeat, sat out one cycle	All	48	83771	0.03	0.0218	0.0134	0.0083	0.02	0.12	-0.03	0.08
Repeat, sat out two+ cycles	All	29	47172	0.04	0.0270	0.0145	0.0125	0.03	0.19	-0.08	0.13
First time	All	54	91735	0.05	0.0228	0.0136	0.0092	0.08	0.13	0.02	0.14
All	1C171	13	7901	0.06	0.0569	0.0320	0.0250	0.06	0.18	-0.12	0.24
All	1P071	14	7639	-0.13	0.0235	0.0328	-0.0093	-0.20	0.00	-0.33	-0.07
All	1W071	13	7293	0.18	0.0702	0.0430	0.0272	0.26	0.22	0.04	0.48
All	2S051	15	16868	0.01	0.0462	0.0238	0.0224	0.03	0.21	-0.14	0.20
All	2S071	14	24263	-0.10	0.0109	0.0076	0.0033	-0.14	0.08	-0.22	-0.06
All	2W151	14	15488	0.21	0.0235	0.0290	-0.0055	0.28	0.00	0.14	0.41
All	2W171	12	20397	0.05	0.0142	0.0084	0.0058	0.06	0.11	-0.05	0.16
All	3E771	14	9628	0.06	0.0104	0.0268	-0.0164	0.07	0.00	-0.01	0.16
All	3P051	12	64370	0.10	0.0087	0.0076	0.0011	0.15	0.04	0.07	0.24
All	3P071	11	53468	-0.02	0.0065	0.0038	0.0028	-0.02	0.07	-0.10	0.06
All	4N051	14	15279	-0.03	0.0370	0.0379	-0.0009	-0.04	0.02	-0.22	0.13
All	4N071	13	17210	0.04	0.0072	0.0098	-0.0026	0.05	0.00	-0.04	0.14

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean rho = sample-size-weighted standardized *r* corrected for unreliability in item sets; sd rho = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 20. Non-Hispanic-Hispanic Differences on PFE Items by Item Use History and PFE

Item Condition	PFE	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	mean_rho	sd_rho	CI_LL_95	CI_UL_95
All	All	40	1187980	.04	.0060	.0008	.0051	.05	.13	.01	.10
Repeat, sat out one cycle	All	9	324183	.05	.0040	.0007	.0033	.08	.09	.00	.16
Repeat, sat out two+ cycles	All	9	324183	.04	.0014	.0007	.0007	.06	.05	.01	.11
First time	All	22	539614	.03	.0103	.0010	.0093	.03	.18	-.05	.12
All	00035	14	488280	.03	.0021	.0012	.0009	.04	.05	.00	.08
All	00036	13	434746	.05	.0026	.0006	.0021	.08	.06	.04	.13
All	00037	13	264954	.02	.0193	.0006	.0186	.03	.20	-.09	.16

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean rho = sample-size-weighted standardized *r* corrected for unreliability in item sets; sd rho = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 21. Correlations between Scores on SKT Items and AFQT Scores by Item Use History and SKT

Item Condition	SKT	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	mean_rho	sd_rho	CI_LL_95	CI_UL_95
All	All	216	254458	0.19	0.0084	0.0008	0.0076	0.28	0.12	0.26	0.29
Repeat, consecutive cycles	All	38	38024	0.22	0.0078	0.0009	0.0069	0.27	0.11	0.24	0.31
Repeat, sat out one cycle	All	64	81276	0.22	0.0091	0.0007	0.0084	0.27	0.11	0.24	0.30
Repeat, sat out two+ cycles	All	37	45782	0.14	0.0052	0.0008	0.0044	0.25	0.13	0.20	0.29
First time	All	77	89376	0.18	0.0073	0.0008	0.0065	0.30	0.13	0.27	0.33
All	1C151	15	10402	0.12	0.0025	0.0014	0.0011	0.17	0.03	0.13	0.20
All	1C171	13	7259	0.09	0.0030	0.0018	0.0013	0.13	0.04	0.08	0.17
All	1P051	14	4959	0.18	0.0050	0.0027	0.0024	0.26	0.08	0.19	0.33
All	1P071	14	6580	0.10	0.0048	0.0021	0.0027	0.15	0.08	0.09	0.21
All	1W051	14	5382	0.24	0.0055	0.0023	0.0032	0.35	0.06	0.30	0.41
All	1W071	13	6565	0.21	0.0039	0.0018	0.0021	0.29	0.06	0.24	0.34
All	2S051	15	14329	0.15	0.0041	0.0010	0.0031	0.22	0.07	0.17	0.26
All	2S071	14	20782	0.08	0.0013	0.0007	0.0006	0.11	0.03	0.08	0.14
All	2W151	14	13459	0.21	0.0023	0.0009	0.0013	0.29	0.06	0.25	0.34
All	2W171	12	18045	0.15	0.0058	0.0006	0.0051	0.21	0.08	0.15	0.26
All	3E751	14	7503	0.29	0.0063	0.0016	0.0047	0.41	0.07	0.36	0.46
All	3E771	14	8510	0.19	0.0051	0.0015	0.0035	0.27	0.07	0.21	0.32
All	3P051	12	54833	0.26	0.0059	0.0002	0.0057	0.38	0.07	0.34	0.43
All	3P071	11	46563	0.18	0.0047	0.0002	0.0045	0.26	0.07	0.21	0.31
All	4N051	14	13648	0.29	0.0105	0.0009	0.0096	0.47	0.07	0.42	0.52
All	4N071	13	15639	0.19	0.0074	0.0008	0.0066	0.32	0.10	0.25	0.38

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean rho = sample-size-weighted standardized *r* corrected for unreliability in item sets; sd rho = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 22. Correlations between Scores on PFE Items and AFQT Scores by Item Use History and PFE

Item Condition	PFE	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	mean_rho	sd_rho	CI_LL_95	CI_UL_95
All	All	40	1039369	.20	.0054	.0000	.0054	.31	.10	.28	.34
Repeat, sat out one cycle	All	9	284624	.22	.0023	.0000	.0023	.32	.05	.28	.35
Repeat, sat out two+ cycles	All	9	284624	.21	.0034	.0000	.0033	.32	.06	.27	.37
First time	All	22	470121	.19	.0087	.0000	.0087	.29	.14	.23	.36
All	00035	14	419351	.23	.0024	.0000	.0024	.34	.06	.31	.38
All	00036	13	384783	.19	.0088	.0000	.0088	.29	.13	.21	.37
All	00037	13	235235	.18	.0046	.0001	.0046	.27	.10	.21	.33

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean rho = sample-size-weighted standardized *r* corrected for unreliability in item sets; sd rho = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 23. Correlations between Scores on SKT Items and Time in Service by Item Use History and SKT

Item Condition	SKT	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	mean_rho	sd_rho	CI_LL_95	CI_UL_95
All	All	216	259792	.01	.0082	.0008	.0073	.02	.12	.00	.04
Repeat, consecutive cycles	All	38	38786	.04	.0083	.0010	.0073	.06	.11	.02	.10
Repeat, sat out one cycle	All	64	83014	.03	.0102	.0008	.0094	.04	.12	.00	.07
Repeat, sat out two+ cycles	All	37	46828	.02	.0063	.0008	.0055	.03	.14	-.02	.08
First time	All	77	91164	-.02	.0060	.0008	.0052	-.03	.12	-.06	.00
All	1C151	15	10432	-.07	.0013	.0014	-.0002	-.10	.00	-.12	-.07
All	1C171	13	7348	.03	.0015	.0018	-.0003	.04	.00	.01	.07
All	1P051	14	5011	-.04	.0045	.0028	.0017	-.06	.07	-.12	.00
All	1P071	14	6957	.08	.0013	.0020	-.0007	.12	.00	.09	.15
All	1W051	14	5406	-.05	.0077	.0026	.0051	-.07	.11	-.14	.01
All	1W071	13	6647	-.01	.0035	.0020	.0016	-.02	.06	-.07	.03
All	2S051	15	14510	.00	.0017	.0010	.0006	.01	.04	-.03	.04
All	2S071	14	21642	.03	.0021	.0006	.0014	.04	.05	.01	.08
All	2W151	14	13579	-.10	.0047	.0010	.0037	-.13	.09	-.19	-.07
All	2W171	12	18619	.02	.0030	.0006	.0024	.03	.06	-.02	.08
All	3E751	14	7523	-.06	.0063	.0018	.0045	-.08	.11	-.15	-.01
All	3E771	14	8724	.06	.0037	.0016	.0021	.09	.06	.04	.14
All	3P051	12	55305	-.04	.0022	.0002	.0020	-.07	.06	-.11	-.03
All	3P071	11	48366	.13	.0075	.0002	.0073	.18	.11	.11	.26
All	4N051	14	13709	-.04	.0019	.0010	.0009	-.06	.05	-.10	-.02
All	4N071	13	16014	.06	.0022	.0008	.0014	.10	.04	.06	.14

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean rho = sample-size-weighted standardized *r* corrected for unreliability in item sets; sd rho = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 24. Correlations between Scores on PFE Items and Time in Service by Item Use History and PFE

Item Condition	PFE	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	mean_rho	sd_rho	CI_LL_95	CI_UL_95
All	All	40	1060819	-.07	.0082	.0000	.0082	-.10	.13	-.14	-.06
Repeat, sat out one cycle	All	9	290635	-.06	.0102	.0000	.0101	-.10	.15	-.21	.02
Repeat, sat out two+ cycles	All	9	290635	-.06	.0103	.0000	.0103	-.10	.15	-.21	.02
First time	All	22	479549	-.07	.0070	.0000	.0069	-.11	.13	-.16	-.05
All	00035	14	422172	-.03	.0009	.0000	.0009	-.04	.04	-.07	-.02
All	00036	13	395193	-.02	.0009	.0000	.0009	-.03	.04	-.05	.00
All	00037	13	243454	-.22	.0034	.0000	.0034	-.32	.09	-.37	-.26

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean rho = sample-size-weighted standardized *r* corrected for unreliability in item sets; sd rho = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 25. Correlations between Scores on SKT Items and Time in Grade by Item Use History and SKT

Item Condition	SKT	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	mean_rho	sd_rho	CI_LL_95	CI_UL_95
All	All	216	259792	-.02	.0041	.0008	.0033	-.02	.08	-.03	-.01
Repeat, consecutive cycles	All	38	38786	.02	.0040	.0010	.0030	.02	.07	.00	.05
Repeat, sat out one cycle	All	64	83014	.00	.0042	.0008	.0035	.00	.07	-.02	.01
Repeat, sat out two+ cycles	All	37	46828	-.02	.0024	.0008	.0016	-.03	.07	-.06	.00
First time	All	77	91164	-.04	.0038	.0008	.0029	-.08	.08	-.10	-.06
All	1C151	15	10432	-.04	.0028	.0014	.0013	-.06	.05	-.10	-.02
All	1C171	13	7348	-.08	.0013	.0017	-.0005	-.11	.00	-.14	-.08
All	1P051	14	5011	-.03	.0045	.0028	.0017	-.04	.06	-.10	.02
All	1P071	14	6957	-.01	.0017	.0020	-.0003	-.01	.00	-.05	.03
All	1W051	14	5406	-.04	.0081	.0026	.0055	-.05	.11	-.13	.02
All	1W071	13	6647	-.13	.0026	.0019	.0008	-.17	.05	-.22	-.12
All	2S051	15	14510	.02	.0022	.0010	.0012	.03	.05	.00	.07
All	2S071	14	21642	.00	.0013	.0006	.0007	.01	.04	-.02	.04
All	2W151	14	13579	-.08	.0059	.0010	.0049	-.11	.10	-.17	-.04
All	2W171	12	18619	-.02	.0012	.0006	.0006	-.03	.03	-.06	.00
All	3E751	14	7523	-.05	.0076	.0019	.0057	-.06	.11	-.13	.02
All	3E771	14	8724	.00	.0030	.0016	.0014	.01	.05	-.04	.05
All	3P051	12	55305	-.04	.0020	.0002	.0018	-.07	.06	-.10	-.03
All	3P071	11	48366	.04	.0029	.0002	.0027	.06	.07	.01	.11
All	4N051	14	13709	-.02	.0017	.0010	.0007	-.04	.04	-.07	.00
All	4N071	13	16014	.01	.0013	.0008	.0004	.03	.03	-.01	.06

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean rho = sample-size-weighted standardized *r* corrected for unreliability in item sets; sd rho = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 26. Correlations between Scores on PFE Items and Time in Grade by Item Use History and PFE

Item Condition	PFE	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	mean_rho	sd_rho	CI_LL_95	CI_UL_95
All	All	40	1060819	.00	.0024	.0000	.0023	.01	.07	-.02	.03
Repeat, sat out one cycle	All	9	290635	.02	.0028	.0000	.0027	.03	.08	-.03	.09
Repeat, sat out two+ cycles	All	9	290635	.02	.0027	.0000	.0027	.02	.08	-.04	.09
First time	All	22	479549	-.01	.0016	.0000	.0016	-.02	.06	-.05	.01
All	00035	14	422172	-.02	.0010	.0000	.0010	-.03	.05	-.05	.00
All	00036	13	395193	.05	.0012	.0000	.0012	.08	.05	.04	.11
All	00037	13	243454	-.04	.0007	.0001	.0006	-.05	.04	-.08	-.03

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; mean rho = sample-size-weighted standardized *r* corrected for unreliability in item sets; sd rho = standard deviation of corrected effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Another noteworthy effect is the strong correlation between TIS and performance on PFE items for Airmen completing the 00037 PFE (final row of Table 24). Longer tenured Airmen tended to have lower performance on PFE items compared with less tenured Airmen, but this effect was noticeably larger on PFE 00037 than PFE 00035 and PFE 00036. PFE grade is a substantive moderator of the correlation between performance on PFE items and TIS, as the resulting confidence interval for 00037 (-.37 to -.26) does not overlap with the confidence intervals for 00035 (-.07 to -.02) and 00036 (-.05 to .00). This might reflect a lower degree of motivation or effort on the PFE among long-tenured Airmen, or it might be associated with a career plateau at the E-6 level where some Airmen do not exhibit the qualifications for promotion to the Senior Non-Commissioned Officer level, regardless of their time-in-service.[8]

3.2 Analysis 2: Item Familiarity Effects

Analysis 2 was also concerned with item exposure effects but with a slight difference from Analysis 1: The focus was on item exposure effects operationally defined by an Airman seeing a specific test item more than once (e.g., if an Airman completed an SKT in 2013 and 2014 and there were items on the 2014 Revision that had appeared on the 2013 Revision).

3.2.1 Method

Sample. The sample used in this study was based on the same Airmen included in Analysis 1. However, for Analysis 2, we restricted the sample to Airmen who were testing for at least the second time within an SKT or PFE level. We also restricted the sample to Airmen and SKT/PFE administrations where we could identify items the Airmen had seen on a previous administration. We did not include any 2011 data in this analysis, because we did not know which specific items on the 2011 exams the Airmen would have seen on a previous administration.

Analyses. For each retesting candidate within an exam administration year (promotion board cycle), each item on each revision of an SKT or PFE was coded according to whether the candidate had (a) seen the item on a previous administration or (b) not seen the item on a previous administration. We merged SKT and PFE item responses for each year (revision) with the corresponding promotion board cycle information (sex, race, ethnicity, AFQT scores, TIS scores, and TIG scores). We combined this information with vectors indicating the "item previously seen" status of each item for each Airman.

For each candidate within each SKT or PFE administration (revision), we computed up to two scores based on the item previously seen statuses. In a given administration, an Airman could have a score based on items the Airman had never seen before (new items and repeat items that were not included on that Airman's previous revisions) and items the Airman had seen before (items that were included on one or more of the Airman's previous revisions). We required Airmen to have at least five items in each condition in any administration to receive a score.

Because the number of items in each condition varied between Airmen (i.e., two Airmen in the same SKT and the same revision would have different items they had seen before), we did not

[8] Although it is not evident from the results presented in table 23, this pattern occurred in the SKT as well. Airmen at grades E-5 and E-6 completed the same SKT versions. When the correlations between TIS and SKT scores are calculated within grade, there is a weak correlation between TIS and SKT scores for E-5 Airmen and a stronger negative correlation between TIS and SKT scores for E-6 Airmen.

have a consistent estimate of score reliability. For this reason, Analysis 2 is based on observed scores only with no correction for measurement error.

Within each revision of each SKT and PFE, we calculated mean scores for each item previously seen status for men and women as well as for the following demographic groups: White, Black, Asian, American Indian or Alaska Native, Hispanic, and Non-Hispanic. We calculated *d* values in scores based on sex, race, and ethnic groups for any condition with greater than 20 individuals in each group. We meta-analyzed those *d* values using the *psychmeta* package for R.

Finally, we computed correlations between scores in each item use condition with AFQT scores, time in service scores, and time in grade scores. We meta-analyzed those correlations in the *psychmeta* package for R and explored the potential moderating effect of items previously seen.

3.2.2 Results

Sex Differences by Item Status. Table 27 displays mean differences in scores on SKT items as a function of items previously seen and sex. The average effects tended to be small to moderate, with an overall sex difference (*d*) of 0.20. The overall effect sizes by SKT ranged from -0.03 (scores for 2S071 women were slightly higher, on average, than scores for 2S071 men) to 0.44 (scores for 2W171 men were moderately higher, on average, than scores for 2W171 women).

Focusing on the results by item seen status, the magnitude of group differences by sex were not moderated by item seen history. The *d* values are very similar across conditions, and the variance in effect sizes within levels of item seen history is comparable to the variance in effect sizes overall. For items not seen before, the mean group difference was 0.18 (SD res[9] = 0.15), which is not much different from the magnitude of the sex difference for items previously seen (0.22, SD res = 0.20).

Table 28 displays mean differences in scores on PFE items as a function of item seen status and sex. The mean effects were very small, with an overall sex difference (*d*) of 0.02. The overall effect sizes by PFE (00035, 00036, and 00037) ranged from -0.02 (00037) to 0.05 (00036).

Focusing on the results by item seen status, the magnitude of the group differences by sex were not moderated by item seen history. The *d* values are very small in each item seen condition, and the variance in effect sizes within levels of item seen history is comparable to the variance in effect sizes overall. For items not seen before, the mean group difference was 0.06 (SD res = 0.09), which is not much different from the magnitude of the group difference for items seen before (-0.01, SD res = 0.07).

Race Differences by Item Seen Status. Tables 29-31 present results for score differences on SKT items by race (White-Black, White-Asian, and White-AI/AN). The largest race differences were associated with White-Black comparisons (Table 29), with an overall White-Black *d* value of 0.25 (SD res = 0.11). The overall *d* comparing average scores for White Airmen and average scores for Asian Airmen was 0.13 (SD res = 0.00), and the White-AI/AN *d* = 0.24 (SD res = 0.00). Item seen status did not moderate the magnitude of group differences. Across item seen conditions, the White-Black and White-Asian *d* values were similar in magnitude, and there was no noticeable reduction in variance within item seen condition.

[9] SD res is the standard deviation in effect sizes not attributable to sampling error and, thus, reflects true variability in the effect across conditions.

Table 27. Sex Differences on SKT Items by Item Seen Status and SKT

Item Condition	SKT	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	104	68162	0.20	0.0417	0.0100	0.0317	0.18	0.16	0.24
Items not seen before	All	52	34081	0.18	0.0336	0.0100	0.0236	0.15	0.13	0.23
Items seen before	All	52	34081	0.22	0.0499	0.0100	0.0399	0.20	0.16	0.28
All	1C151	6	2452	0.04	0.0106	0.0206	-0.0100	0.00	-0.07	0.15
All	1C171	8	2742	0.05	0.0137	0.0186	-0.0049	0.00	-0.05	0.15
All	1P051	4	866	0.14	0.0170	0.0361	-0.0191	0.00	-0.07	0.34
All	1P071	8	2444	0.12	0.0165	0.0184	-0.0019	0.00	0.01	0.23
All	1W051	6	904	0.15	0.0272	0.0369	-0.0097	0.00	-0.02	0.32
All	1W071	8	2554	0.33	0.0409	0.0210	0.0199	0.14	0.16	0.50
All	2S051	8	3534	-0.01	0.0063	0.0102	-0.0039	0.00	-0.08	0.06
All	2S071	8	8140	-0.03	0.0049	0.0040	0.0008	0.03	-0.08	0.03
All	2W151	8	3010	0.40	0.0171	0.0205	-0.0034	0.00	0.30	0.51
All	2W171	8	6838	0.44	0.0469	0.0155	0.0314	0.18	0.26	0.63
All	3P051	8	9790	0.21	0.0069	0.0051	0.0018	0.04	0.14	0.28
All	3P071	8	15424	0.34	0.0212	0.0044	0.0168	0.13	0.22	0.47
All	4N051	8	3198	0.07	0.0071	0.0105	-0.0034	0.00	0.00	0.14
All	4N071	8	6266	0.05	0.0088	0.0052	0.0036	0.06	-0.03	0.13

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 28. Sex Differences on PFE Items by Item Seen Status and PFE

Item Condition	PFE	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	18	215992	0.02	0.0078	0.0005	0.0072	0.09	-0.02	0.07
Items not seen before	All	9	107996	0.06	0.0082	0.0005	0.0077	0.09	-0.01	0.13
Items seen before	All	9	107996	-0.01	0.0054	0.0005	0.0049	0.07	-0.07	0.05
All	00035	6	52402	0.02	0.0026	0.0007	0.0019	0.04	-0.03	0.07
All	00036	6	108982	0.05	0.0047	0.0003	0.0043	0.07	-0.02	0.12
All	00037	6	54608	-0.02	0.0183	0.0007	0.0177	0.13	-0.17	0.12

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 29. White-Black Differences on SKT Items by Item Seen Status and SKT

Item Condition	SKT	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	100	59392	0.25	0.0224	0.0100	0.0124	0.11	0.22	0.28
Items not seen before	All	50	29696	0.23	0.0216	0.0100	0.0117	0.11	0.18	0.27
Items seen before	All	50	29696	0.28	0.0220	0.0100	0.0121	0.11	0.24	0.32
All	1C151	6	2208	0.12	0.0445	0.0243	0.0202	0.14	-0.10	0.34
All	1C171	4	1456	0.29	0.0053	0.0192	-0.0139	0.00	0.17	0.40
All	1P071	8	2094	0.21	0.0092	0.0186	-0.0094	0.00	0.13	0.29
All	1W071	4	1328	0.36	0.0411	0.0331	0.0081	0.09	0.04	0.68
All	2S051	8	2992	0.13	0.0262	0.0111	0.0150	0.12	-0.01	0.26
All	2S071	8	6484	0.06	0.0020	0.0050	-0.0030	0.00	0.02	0.09
All	2W151	8	2608	0.25	0.0218	0.0177	0.0041	0.06	0.13	0.37
All	2W171	8	5684	0.32	0.0273	0.0080	0.0193	0.14	0.18	0.46
All	3E751	6	1528	0.39	0.0077	0.0278	-0.0202	0.00	0.30	0.49
All	3E771	8	2894	0.39	0.0220	0.0216	0.0004	0.02	0.27	0.51
All	3P051	8	8754	0.26	0.0071	0.0046	0.0026	0.05	0.19	0.33
All	3P071	8	13620	0.32	0.0104	0.0035	0.0069	0.08	0.23	0.40
All	4N051	8	2716	0.24	0.0125	0.0149	-0.0024	0.00	0.15	0.34
All	4N071	8	5026	0.23	0.0223	0.0076	0.0147	0.12	0.11	0.36

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 30. White-Asian Differences on SKT Items by Item Seen Status and SKT

Item Condition	SKT	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	48	31976	0.13	0.0250	0.0251	-0.0001	0.00	0.09	0.18
Items not seen before	All	24	15988	0.13	0.0236	0.0251	-0.0016	0.00	0.06	0.19
Items seen before	All	24	15988	0.14	0.0275	0.0251	0.0024	0.05	0.07	0.21
All	2S051	6	1810	0.00	0.0142	0.0368	-0.0226	0.00	-0.12	0.13
All	2S071	8	3456	-0.13	0.0274	0.0190	0.0083	0.09	-0.26	0.01
All	2W151	2	846	0.04	0.0406	0.0444	-0.0038	0.00	-1.77	1.85
All	2W171	8	4692	0.16	0.0197	0.0268	-0.0070	0.00	0.04	0.28
All	3P051	4	4606	0.17	0.0036	0.0236	-0.0200	0.00	0.07	0.26
All	3P071	8	10988	0.17	0.0126	0.0227	-0.0102	0.00	0.07	0.26
All	4N051	6	2056	0.24	0.0277	0.0394	-0.0116	0.00	0.06	0.41
All	4N071	6	3522	0.23	0.0181	0.0195	-0.0013	0.00	0.09	0.37

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 31. White-AI/AN Differences on SKT Items by Item Seen Status and SKT

Item Condition	SKT	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	8	12118	0.24	0.0396	0.0421	-0.0025	0.00	0.07	0.40
Items not seen before	All	4	6059	0.16	0.0442	0.0421	0.0021	0.05	-0.17	0.50
Items seen before	All	4	6059	0.31	0.0343	0.0421	-0.0079	0.00	0.01	0.60
All	3P051	4	5002	0.37	0.0465	0.0432	0.0033	0.06	0.03	0.72
All	3P071	4	7116	0.14	0.0160	0.0414	-0.0254	0.00	-0.06	0.34

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

As occurred with Analysis 1, the White-AI/AN comparisons (Table 31) revealed differences in effect sizes across item seen statuses, with smaller group differences associated with items not seen before and larger group differences associated with items previously seen. However, the numbers of conditions and total sample sizes for the White-AI/AN comparisons were much smaller than other analyses, and the confidence intervals overlapped substantially. As such, item seen history does not seem to be a true moderator of the White-AI/AN group difference.

Tables 32-34 present group difference results for the PFE. As with the SKT results, the group differences were generally small and were not moderated by item seen history.

Ethnicity Differences by Item Status. Tables 35 and 36 present group difference results on SKT and PFE items by ethnicity (Hispanic or non-Hispanic). The differences were very small, with *d* values near zero. The magnitude of the *d* values was similar across item seen conditions, and the standard deviations within item conditions were not reduced compared with the overall standard deviations.

AFQT-Performance Correlations by Item Status. Tables 37 and 38 display the meta-analysis results for the AFQT score correlations for SKT and PFE items, respectively. Scores across SKT and PFE items were correlated with AFQT scores, with an weighted average AFQT-SKT correlation of .20 (SD = .07) and a weighted average AFQT-PFE correlation of .22 (SD = .05). Item seen history did not act as a substantive moderator, with nearly equal *r* values across conditions and standard deviations as large within item use condition as they were overall.

Table 32. White-Black Differences on PFE Items by Item Seen Status and PFE

Item Condition	PFE	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	18	185190	0.11	0.0061	0.0006	0.0055	0.07	0.08	0.15
Items not seen before	All	9	92595	0.14	0.0055	0.0006	0.0049	0.07	0.08	0.20
Items seen before	All	9	92595	0.09	0.0060	0.0006	0.0054	0.07	0.03	0.15
All	00035	6	46240	0.05	0.0008	0.0007	0.0001	0.01	0.02	0.08
All	00036	6	93724	0.17	0.0016	0.0004	0.0012	0.03	0.13	0.21
All	00037	6	45226	0.07	0.0085	0.0008	0.0077	0.09	-0.03	0.17

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 33. White-Asian Differences on PFE Items by Item Seen Status and PFE

Item Condition	PFE	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	18	153354	0.10	0.0083	0.0023	0.0060	0.08	0.05	0.14
Items not seen before	All	9	76677	0.07	0.0057	0.0023	0.0034	0.06	0.01	0.13
Items seen before	All	9	76677	0.12	0.0106	0.0023	0.0082	0.09	0.04	0.20
All	00035	6	37446	-0.02	0.0030	0.0033	-0.0003	0.00	-0.07	0.04
All	00036	6	78736	0.13	0.0052	0.0014	0.0038	0.06	0.05	0.20
All	00037	6	37172	0.14	0.0041	0.0034	0.0007	0.03	0.08	0.21

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 34. White-AI/AN Differences on PFE Items by Item Seen Status and PFE

Item Condition	PFE	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	18	147248	0.10	0.0141	0.0088	0.0054	0.07	0.04	0.16
Items not seen before	All	9	73624	0.11	0.0124	0.0088	0.0037	0.06	0.02	0.19
Items seen before	All	9	73624	0.10	0.0176	0.0088	0.0088	0.09	0.00	0.20
All	00035	6	36106	0.00	0.0074	0.0102	-0.0027	0.00	-0.09	0.09
All	00036	6	75450	0.13	0.0061	0.0053	0.0008	0.03	0.04	0.21
All	00037	6	35692	0.16	0.0275	0.0146	0.0129	0.11	-0.01	0.34

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 35. Non-Hispanic-Hispanic Differences on SKT Items by Item Seen Status and SKT

item Condition	SKT	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	68	59300	0.01	0.0290	0.0187	0.0102	0.10	-0.03	0.05
Items not seen before	All	34	29650	0.01	0.0281	0.0187	0.0093	0.10	-0.05	0.07
Items seen before	All	34	29650	0.01	0.0307	0.0187	0.0120	0.11	-0.05	0.07
All	1C171	4	1722	-0.05	0.0095	0.0343	-0.0248	0.00	-0.21	0.10
All	1P071	6	2142	-0.07	0.0449	0.0411	0.0038	0.06	-0.30	0.15
All	1W071	6	2314	0.33	0.0671	0.0494	0.0178	0.13	0.06	0.61
All	2S051	4	2350	-0.13	0.0471	0.0370	0.0101	0.10	-0.47	0.22
All	2S071	8	8140	-0.12	0.0194	0.0112	0.0082	0.09	-0.24	-0.01
All	2W151	4	2272	0.25	0.0785	0.0398	0.0387	0.20	-0.19	0.70
All	2W171	8	6838	0.03	0.0145	0.0136	0.0009	0.03	-0.07	0.13
All	3E771	6	2996	0.08	0.0108	0.0323	-0.0215	0.00	-0.03	0.19
All	3P051	6	8836	-0.03	0.0187	0.0206	-0.0019	0.00	-0.17	0.11
All	3P071	8	15424	-0.01	0.0126	0.0080	0.0046	0.07	-0.10	0.09
All	4N071	8	6266	0.09	0.0051	0.0139	-0.0088	0.00	0.03	0.15

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 36. Non-Hispanic-Hispanic Differences on PFE Items by Item Seen Status and PFE

Item Condition	PFE	*k*	*N*	mean_*d*	var_*d*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	18	215992	0.08	0.0053	0.0016	0.0037	0.06	0.04	0.11
Items not seen before	All	9	107996	0.09	0.0045	0.0016	0.0029	0.05	0.04	0.14
Items seen before	All	9	107996	0.06	0.0065	0.0016	0.0049	0.07	0.00	0.13
All	00035	6	52402	0.02	0.0048	0.0034	0.0015	0.04	-0.05	0.10
All	00036	6	108982	0.13	0.0011	0.0009	0.0002	0.01	0.09	0.16
All	00037	6	54608	0.02	0.0035	0.0012	0.0023	0.05	-0.04	0.08

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *d* = sample-size-weighted standardized mean difference between groups; var *d* = variance of sample-size-weighted *d*; var *e* = variance in *d* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 37. Correlations between Scores on SKT Items and AFQT Scores by Item Seen Status and SKT

Item Condition	SKT	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	128	70658	0.20	0.0061	0.0017	0.0044	0.07	0.19	0.21
Items not seen before	All	64	35329	0.18	0.0051	0.0017	0.0034	0.06	0.16	0.20
Items seen before	All	64	35329	0.22	0.0066	0.0017	0.0049	0.07	0.19	0.24
All	1C151	8	2556	0.09	0.0013	0.0031	-0.0018	0.00	0.06	0.13
All	1C171	8	2666	0.11	0.0069	0.0029	0.0040	0.06	0.04	0.18
All	1P051	8	1078	0.20	0.0041	0.0069	-0.0028	0.00	0.15	0.25
All	1P071	8	2280	0.16	0.0055	0.0033	0.0022	0.05	0.10	0.23
All	1W051	8	950	0.21	0.0035	0.0078	-0.0043	0.00	0.16	0.26
All	1W071	8	2514	0.25	0.0027	0.0028	-0.0001	0.00	0.21	0.29
All	2S051	8	3408	0.17	0.0044	0.0022	0.0022	0.05	0.11	0.22
All	2S071	8	7728	0.12	0.0012	0.0010	0.0002	0.01	0.09	0.15
All	2W151	8	2864	0.21	0.0062	0.0026	0.0037	0.06	0.14	0.27
All	2W171	8	6548	0.18	0.0085	0.0011	0.0073	0.09	0.10	0.25
All	3E751	8	1808	0.30	0.0014	0.0037	-0.0023	0.00	0.27	0.33
All	3E771	8	3186	0.28	0.0012	0.0021	-0.0009	0.00	0.25	0.31
All	3P051	8	9324	0.20	0.0004	0.0008	-0.0004	0.00	0.18	0.21
All	3P071	8	14562	0.23	0.0041	0.0005	0.0036	0.06	0.18	0.29
All	4N051	8	3116	0.25	0.0063	0.0023	0.0040	0.06	0.18	0.32
All	4N071	8	6070	0.24	0.0060	0.0012	0.0049	0.07	0.18	0.31

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 38. Correlations between Scores on PFE Items and AFQT Scores by Item Seen Status and PFE

Item Condition	PFE	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	18	204014	0.22	0.0026	0.0001	0.0025	0.05	0.20	0.25
Items not seen before	All	9	102007	0.25	0.0011	0.0001	0.0010	0.03	0.23	0.28
Items seen before	All	9	102007	0.19	0.0025	0.0001	0.0024	0.05	0.16	0.23
All	00035	6	49592	0.18	0.0025	0.0001	0.0024	0.05	0.13	0.23
All	00036	6	102618	0.25	0.0017	0.0001	0.0016	0.04	0.20	0.29
All	00037	6	51804	0.22	0.0025	0.0001	0.0024	0.05	0.17	0.27

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Tenure-Performance Correlations by Item Status. Tables 39-42 present results of correlations between tenure (TIS, TIG) and scores on SKT and PFE items. The correlation between performance on SKT items and TIS was small, with an overall weighted average correlation of -.04 (SD = .14). The correlation between performance on PFE items and TIS was stronger and negative, with an overall weighted average correlation of -.19 (SD = .12). Within item seen conditions, the correlations between scores and time-in-service were consistent. Further, as with Analysis 1, there was a stronger negative relation between time-in-service and performance on PFE items for PFE 00037.

The correlation between TIG and performance on SKT items (-.14) was stronger than the correlation between TIS and performance on SKT. There was also some evidence of moderation by item seen status: Airmen were less likely to answer SKT items correctly if they had been in the same grade for a long time, and this effect was stronger on items they had not seen before (rows 2 and 3 of table 41). Item seen status did not moderate the strength of the relation between TIG and performance on PFE items.

3.2.3 Discussion

The results from Analysis 2 are consistent in suggesting that item seen status does not moderate relations between performance on items and demographic and experience factors. The magnitude of *d* values by sex, race, and ethnicity were consistent across item seen history. Correlations between performance and AFQT and tenure also tended to be consistent across item seen status.

As with Analysis 1, the observation that stood out in Analysis 2 was the strong correlation between TIS and performance on PFE items for PFE 00037. As suggested earlier, this pattern is consistent with a type of ceiling effect among Airmen who make it to E-6, but are not promoted to E-7.

Table 39. Correlations between Scores on SKT Items and Time in Service by Item Seen Status and SKT

Item Condition	SKT	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	128	72562	-0.04	0.0215	0.0018	0.0198	0.14	-0.06	-0.01
Items not seen before	All	64	36281	-0.06	0.0250	0.0018	0.0232	0.15	-0.10	-0.02
Items seen before	All	64	36281	-0.01	0.0175	0.0018	0.0158	0.13	-0.05	0.02
All	1C151	8	2564	-0.14	0.0210	0.0030	0.0179	0.13	-0.26	-0.02
All	1C171	8	2698	-0.02	0.0060	0.0030	0.0030	0.05	-0.09	0.04
All	1P051	8	1098	-0.21	0.0072	0.0068	0.0005	0.02	-0.28	-0.13
All	1P071	8	2430	-0.03	0.0110	0.0033	0.0077	0.09	-0.12	0.06
All	1W051	8	952	-0.06	0.0335	0.0085	0.0250	0.16	-0.21	0.09
All	1W071	8	2538	-0.10	0.0030	0.0031	-0.0002	0.00	-0.14	-0.05
All	2S051	8	3460	-0.18	0.0044	0.0022	0.0023	0.05	-0.23	-0.12
All	2S071	8	8082	-0.04	0.0043	0.0010	0.0033	0.06	-0.10	0.01
All	2W151	8	2910	-0.25	0.0038	0.0024	0.0013	0.04	-0.30	-0.19
All	2W171	8	6778	-0.01	0.0064	0.0012	0.0053	0.07	-0.07	0.06
All	3E751	8	1812	-0.25	0.0156	0.0039	0.0117	0.11	-0.35	-0.14
All	3E771	8	3276	-0.01	0.0019	0.0024	-0.0005	0.00	-0.05	0.03
All	3P051	8	9448	-0.05	0.0584	0.0008	0.0576	0.24	-0.25	0.15
All	3P071	8	15194	0.10	0.0035	0.0005	0.0030	0.05	0.05	0.15
All	4N051	8	3122	-0.14	0.0145	0.0025	0.0121	0.11	-0.24	-0.04
All	4N071	8	6200	0.01	0.0074	0.0013	0.0061	0.08	-0.06	0.08

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 40. Correlations between Scores on PFE Items and Time in Service by Item Seen Status and PFE

Item Condition	PFE	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	18	209618	-0.19	0.0139	0.0001	0.0138	0.12	-0.25	-0.13
Items not seen before	All	9	104809	-0.22	0.0103	0.0001	0.0102	0.10	-0.29	-0.14
Items seen before	All	9	104809	-0.16	0.0176	0.0001	0.0175	0.13	-0.27	-0.06
All	00035	6	50198	-0.08	0.0054	0.0001	0.0053	0.07	-0.16	0.00
All	00036	6	105824	-0.16	0.0024	0.0001	0.0023	0.05	-0.21	-0.11
All	00037	6	53596	-0.36	0.0025	0.0001	0.0024	0.05	-0.41	-0.30

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 41. Correlations between Scores on SKT Items and Time in Grade by Item Seen Status and SKT

Item Condition	SKT	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	128	72562	-0.14	0.0204	0.0017	0.0187	0.14	-0.17	-0.12
Items not seen before	All	64	36281	-0.19	0.0236	0.0016	0.0220	0.15	-0.23	-0.16
Items seen before	All	64	36281	-0.09	0.0123	0.0017	0.0105	0.10	-0.12	-0.07
All	1C151	8	2564	-0.10	0.0258	0.0031	0.0227	0.15	-0.24	0.03
All	1C171	8	2698	-0.22	0.0157	0.0027	0.0130	0.11	-0.33	-0.12
All	1P051	8	1098	-0.19	0.0094	0.0068	0.0025	0.05	-0.27	-0.11
All	1P071	8	2430	-0.22	0.0370	0.0030	0.0340	0.18	-0.38	-0.06
All	1W051	8	952	-0.03	0.0268	0.0085	0.0183	0.14	-0.17	0.11
All	1W071	8	2538	-0.34	0.0075	0.0025	0.0050	0.07	-0.41	-0.27
All	2S051	8	3460	-0.18	0.0047	0.0022	0.0026	0.05	-0.23	-0.12
All	2S071	8	8082	-0.17	0.0105	0.0009	0.0096	0.10	-0.26	-0.09
All	2W151	8	2910	-0.25	0.0077	0.0024	0.0053	0.07	-0.32	-0.18
All	2W171	8	6778	-0.17	0.0121	0.0011	0.0110	0.10	-0.26	-0.08
All	3E751	8	1812	-0.25	0.0138	0.0039	0.0099	0.10	-0.35	-0.15
All	3E771	8	3276	-0.18	0.0047	0.0023	0.0024	0.05	-0.24	-0.12
All	3P051	8	9448	-0.08	0.0663	0.0008	0.0655	0.26	-0.29	0.14
All	3P071	8	15194	-0.06	0.0039	0.0005	0.0034	0.06	-0.11	-0.01
All	4N051	8	3122	-0.15	0.0149	0.0025	0.0124	0.11	-0.25	-0.04
All	4N071	8	6200	-0.14	0.0050	0.0012	0.0038	0.06	-0.20	-0.08

Note. SKT = Specialty Knowledge Test; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

Table 42. Correlations between Scores on PFE Items and Time in Grade by Item Seen Status and PFE

Item Condition	PFE	*k*	*N*	mean_*r*	var_*r*	var_*e*	var_res	sd_res	CI_LL_95	CI_UL_95
All	All	18	209618	-0.12	0.0059	0.0001	0.0058	0.08	-0.16	-0.08
Items not seen before	All	9	104809	-0.14	0.0015	0.0001	0.0014	0.04	-0.17	-0.11
Items seen before	All	9	104809	-0.10	0.0103	0.0001	0.0102	0.10	-0.18	-0.02
All	00035	6	50198	-0.08	0.0053	0.0001	0.0052	0.07	-0.16	-0.01
All	00036	6	105824	-0.10	0.0026	0.0001	0.0026	0.05	-0.15	-0.04
All	00037	6	53596	-0.20	0.0057	0.0001	0.0056	0.07	-0.28	-0.12

Note. PFE = Promotion Fitness Examination; *k* = number of effect sizes included in the meta-analysis; *N* = total sample size across all effect sizes; mean *r* = sample-size-weighted correlation; var *r* = variance of sample-size-weighted *r*; var *e* = variance in *r* attributable to sampling error; var res = variance in effect sizes not attributable to sampling error; sd res = standard deviation in effect sizes not attributable to sampling error; CI_LL_95 = lower limit of 95% confidence interval; CI_UL_95 = upper limit of 95% confidence interval.

3.3 Analysis 3: Comparisons of Item-Level Statistics for First-Time and Repeat Examinees

3.3.1 Method

Sample. The sample used in this study consisted of Airmen who (a) were seeking promotion to grade E-5, E-6, or E-7; (b) took a relevant promotion test (a PFE and/or an SKT associated with one of the following AFSCs: 1C1X1, 1P0X1, 1W0X1, 2S0X1, 2W1X1, 3E7X1, 3P0X1, or 4N0X1) for the grade above their current grade; and (c) tested in the 2011 through 2016 promotion cycles. For data-quality assurance, we excluded examinees with incomplete data. We merged Airmen's item-level data with their prior testing records to identify those who had previously taken the exam for which they were sitting. We also merged Airmen's item responses with their AFQT scores; if multiple AFQT scores were on record for a given Airmen, we used the mean of their unique scores as their AFQT score in our analyses. We used items' usage histories to identify which items in each test administration were new and which had been used in at least one previous administration.

Item statistics. For each administration, we computed item statistics using data from first-time examinees and from repeat examinees. The item statistics of interest were item difficulties reflecting the proportion of examinees who got each item correct (i.e., "*p*-values") and ability discrimination indices indicating how examinees' item-level scores correlate with their ability levels. One of the most common methods to estimate discrimination indices is simply to compute correlations between item scores and total scores. However, this approach results in overestimates of ability discrimination, because examinees' scores on the item being examined are included in their total scores. The simplest way to avoid this overestimation is to correct the item-total correlations for spuriousness by omitting the focal item from the total score. Another way to avoid spuriously high discrimination indices is to obtain ability indicators from a source independent of the test in question. In this case, examinees' AFQT scores were available and could provide useful information about how item responses relate to cognitive ability. We used corrected item-total correlations and item-AFQT correlations as indicators of ability discrimination in our analyses. Both indices were computed as point-biserial correlations.

Statistical corrections. Differences in the rates of correct item responses between first-time and repeat examinees were the exclusive focus of our *p*-value analyses. However, differences in *p*-values between groups can also artifactually influence the magnitudes of differences between ability discrimination indices, either by obscuring real differences or creating the illusion of differences where none really exist. Differences between groups' point-biserial correlations are clearly interpretable only when an item has equal difficulty for both groups, which means that the variance of item scores is equal between groups. To rule out differences in item difficulty as an explanation for our results, we corrected all items' point-biserial correlations for variation in *p*-values by adjusting the correlations to what they would have been if 50% of examinees had answered correctly (i.e., the proportion correct at which the dichotomous "correct-incorrect" item score has maximum variance).

Comparisons between ability discrimination indices are further aided if the groups also have equal variance on the ability variable (e.g., the total score in the item-total correlation), as differential variability in ability can also affect the magnitudes of within-group correlations. Selection effects, including those introduced by promoting people from one grade to the next, reduce the variation

among those who remain in the sample after selection decisions have been made. In samples of PFE and SKT examinees, repeat examinees will be less representative of Airmen in their grade because they have previously been passed over for promotion to the next grade. First-time examinees, however, will be more representative of their grade by virtue of having not previously been considered for promotion to the next grade. The potential for differential variability between first-time and repeat examinees should be factored into comparisons of items' ability discrimination to avoid artifactual results. For example, if observed item-total correlations differ between the two examinee groups, the difference could be due to reduced variation in the repeat examinee group as a result of selection/promotion effects. In addition to our corrections for variation in item difficulty, we corrected all correlations for range-variation between groups by adjusting each correlation to what it would have been if both groups had equal variance in ability. Our corrections for differential variability in both item scores and ability scores between first-time and repeat examinees allowed purer indications of the differences in ability discrimination between the two groups.

Mixed-effects models. We summarized the differences ***between*** item statistics for first-time and repeat examinees using mixed-effects linear models computed using the *lme4* package in R. A basic assumption of standard regression models (e.g., OLS linear regression) is that model errors or residuals are independent of one another across observations (e.g., Cohen, Cohen, Aiken, & West, 2003). A benefit of mixed-effects models is that they can account for sources of dependency when estimating the effect of one or more predictor variables on an outcome, which was relevant in the present analyses because item-level observations were not independent of each other in our data.

The dependent variables in our models were constructed by subtracting first-time examinees' item statistics from repeat examinees' item statistics. When the dependent variable in a mixed-effects model is a vector of statistical estimates, the estimates can be weighted so that the model functions as a mixed-effects meta-analysis. It is critical that studies are weighted properly, as poorly chosen weights can lead to invalid model results. Below we describe the weights we used and why they were appropriate for our models.

Weighting method. We weighted estimates as a function of "effective sample size," a random-effects weighting method for group-difference effects that estimates each effect's statistical precision by factoring in both (a) the combined sample size of the groups and (b) the proportional difference in sample size between the groups. For difference statistics, effective sample size can be computed as

$$N_{effective} = 4Np_1(1 - p_1)$$

where N is the total sample size, p_1 is the proportion of the sample that belongs to one of the two groups, and $N_{effective} \leq N$. This equation is related to the assumptions made in the error-variance formula for standardized mean difference statistics; it downwardly adjusts the total sample size when the group-membership variable deviates from a 50/50 split. The variance of a dichotomous variable is computed as $p_1(1 - p_1)$, which takes on a maximum value of .25 when p_1 = .50. Maximizing the variance in group membership is desirable for difference analyses, because difference estimates are more precise when the group-specific values being compared have equal precision. When computing $N_{effective}$, multiplying N by $p_1(1 - p_1)$ imposes a penalty for departures from maximal variance in group membership, because $N_{effective}$ decreases as p_1

departs from .50; multiplying that result by 4 (the reciprocal of .25) puts the result back into the sample-size metric and accounts for the fact that $p_1(1-p_1) \leq .25$. This ensures that $N_{effective}$ equals N only when groups are equally represented.

As an example of how $N_{effective}$ indexes statistical precision, consider two samples of equal size that each include examines who are distributed differently between two groups. Comparisons between the groups will be made with differential precision in each of these samples because of differences in subgroup proportions. If Sample A has 100 examinees from each of two groups (equal representation) and Sample B has 75 examinees from one group and 125 from the other, Sample A's $N_{effective}$ is 200 and Sample B's $N_{effective}$ is 187.5 even though both have a total N of 200. Sample B's departures from maximal group-membership variance reduce the precision of its estimates relative to Sample A's estimates.

Weighting estimates of differences by $N_{effective}$ in a meta-analysis accounts for differences in precision across samples that are not reflected by total sample size alone. $N_{effective}$ weights are also preferable to inverse sampling-variance weights in meta-analyses, because sampling-variance estimates are dependent on the statistical values being meta-analyzed, whereas $N_{effective}$ weights are indices of random-effects precision that are statistically independent of the observations. Recent simulation studies have reported that $N_{effective}$ weights yield less biased estimates than do variance-based weighting schemes (Bakbergenuly, Hoaglin, & Kulinskaya, 2019a; 2019b).

The weights used in our summary models also accounted for the statistical corrections we applied to ability discrimination indices, as applying a statistical correction introduces additional sources of uncertainty and alters the precision of the adjusted statistic. We accounted for this additional uncertainly by adjusting subgroup sample sizes as a function of the magnitudes of the statistical corrections applied to each group's item-ability correlation. Corrections applied to correlations create corresponding changes to the correlations' precision: as the magnitude of the correction increases, the precision of the corrected estimate decreases. This change in precision can be reflected in the sample-size metric by multiplying the sample size by the squared ratio of the observed correlation to the corrected correlation (this ratio is commonly referred to as an attenuation factor in psychometric meta-analysis):

$$n' = n\left(\frac{r}{r_c}\right)^2$$

When applying this formula, the correction-adjusted sample size decreases as the magnitude of the correction increases. We used this method to adjust the sample sizes for first-time and repeat examinees prior to determining the $N_{effective}$ weights for differences in corrected correlations, which allowed the weights to reflect the cumulative precision of each difference estimate after all adjustments were made.

Accounting for dependence among observations. We needed to account for two critical sources of dependence in our models: common samples of examinees and reusage of items. Items grouped together on the same form of a test and administered to the same sample of examinees are dependent by virtue of their item statistics being based on responses from the same set of people. Observations are also dependent when the same item is administered to different samples. We clustered observations by both test administration ID (an identifier we constructed for each unique version of a

test) and item ID in our models, which allowed the dependencies among observations to be directly modeled so that they did not bias the estimation of substantively important effects.

Overview of models. We ran three models for each item statistic, each of which summarizes the differences between first-time and repeat examinees at different levels of aggregation:

- Model 1a: Our simplest model examined differences at the highest level of aggregation. This was a random-intercepts model that summarized the overall difference between first-time and repeat examinees across all administrations of all tests.
- Model 2a: This model built on Model 1a by adding test type (i.e., PFE or SKT) as an independent variable so that separate intercepts (estimates of overall differences between repeat and first-time examinees) could be computed for PFE tests and SKT tests.
- Model 3a: This model built on Model 1a by adding AFSC as an independent variable so that separate intercepts (estimates of overall differences between repeat and first-time examinees) could be computed for each PFE and SKT test.

Magnitudes of differences between item statistics for first-time and repeat examinees could differ as a function of item exposure. For example, items that have been previously exposed through inclusion on an earlier version of an exam may be easier or less effective at discriminating low- from high-ability examinees due to examinees' awareness of the items. This effect might differ between first-time and repeat examinees. We explored this possibility in a set of follow-up analyses in which we added item-exposure status as a moderator to the main-effect models described above. Our supplemental item-exposure models were

- Model 1b: This model builds on Model 1a by testing whether item-exposure status moderates differences between first-time and repeat testers.
- Model 2b: This model builds on Model 2a by separately testing the moderating effect of item-exposure status for PFE tests and SKT tests.
- Model 3b: This model builds on Model 3a by separately testing the moderating effect of item-exposure status for each PFE and SKT test.

3.3.2 Results

Item difficulty. Table 43 displays average differences in item difficulties from linear summary models. The average effects tended to be small, with most differences not exceeding .02 in absolute value. The overall difference in item difficulty between repeat and first-time examinees across all tests was statistically significant but negligible in magnitude (-.02, indicating items are easier on average for first-time examinees), and aggregate differences were also small when estimated separately for PFE tests (-.01) and SKT tests (-.02). At the AFS level, the largest significant differences were -.17 for 3P071 and -.08 for 2S071. These differences are non-trivial in magnitude and are based on large sample sizes for both first-time and repeat examinees. Differences between new and exposed items were virtually non-existent and did not aid interpretations of the aggregate effects. Results for these analyses are reported in Appendix B (see Table B1).

Table 43. Linear Summary Model of Differences between Item Difficulties (p Values) for First-Time Examinees and Repeat Examinees

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	SE	p	95% CILL	95% CIUL
Model 1a: Overall Difference Effect										
Intercept	83	8,260	570,561	240,711	329,850	-0.02	0.01	.01	-0.03	-0.01
Model 2a: Difference Effect by Test Type										
PFE	20	1,998	478,559	205,138	273,421	-0.01	0.01	.49	-0.04	0.02
SKT	63	6,262	92,002	35,573	56,429	-0.02	0.01	.00	-0.04	-0.01
Model 3a: Difference Effect by PFE and SKT Test										
00035	8	800	222,491	126,956	95,535	-0.01	0.02	.71	-0.05	0.03
00036	7	698	173,263	62,902	110,361	-0.01	0.02	.70	-0.05	0.04
00037	5	500	82,805	15,280	67,525	-0.02	0.03	.56	-0.07	0.04
1C151	5	500	3,703	1,790	1,913	0.00	0.03	.87	-0.06	0.05
1C171	2	199	1,153	404	749	-0.01	0.04	.86	-0.09	0.08
1P051	5	499	2,051	945	1,106	0.00	0.03	.93	-0.05	0.05
1P071	2	187	1,046	339	707	-0.01	0.04	.88	-0.09	0.08
1W051	5	500	2,129	1,156	973	-0.02	0.03	.56	-0.07	0.04
1W071	4	388	2,202	493	1,709	-0.02	0.03	.49	-0.08	0.04
2S051	5	500	5,665	2,614	3,051	0.01	0.03	.59	-0.04	0.07
2S071	3	298	5,154	1,140	4,014	-0.08	0.03	.02	-0.15	-0.01
2W151	5	500	5,491	2,651	2,840	-0.01	0.03	.58	-0.07	0.04
2W171	3	300	4,778	1,050	3,728	-0.07	0.03	.04	-0.14	-0.01
3E751	5	500	3,087	1,500	1,587	0.00	0.03	.99	-0.05	0.05
3E771	2	195	1,212	415	797	-0.03	0.04	.44	-0.12	0.05
3P051	5	500	26,951	13,871	13,080	-0.01	0.03	.77	-0.06	0.04
3P071	4	397	18,311	3,322	14,989	-0.17	0.03	.00	-0.22	-0.11
4N051	5	500	5,414	2,958	2,456	0.00	0.03	.90	-0.06	0.05
4N071	3	299	3,655	925	2,730	-0.01	0.03	.77	-0.08	0.06

Note. Term = Label for effect tested in the regression model; k_{Admin} = Number of test administrations that contributed to the effect; k_{Item} = Number of items that contributed to the effect; N = Total size of examinee sample; $n_{Firsttime}$ = Size of first-time examinee sample; n_{Repeat} = Size of repeat examinee sample; b = Regression coefficient indicating the average effect at a given level of aggregation; SE = Standard error of b; $p = p$ value of for the significance test for b; 95% CILL = Lower bound of the 95% confidence interval for b; 95% CIUL = Upper bound of the 95% confidence interval for b. A positive b coefficient indicates that the average item statistic for repeat examinees was larger than for first-time examinees.

The prevalence of negative effects is notable. As mentioned above, negative differences indicate that first-time examinees answered items correctly at higher rates than did repeat examinees. It is important to note that these differences in item-level scores do not control for differences in ability between the groups, so the results must be interpreted cautiously. Repeat examinees have previously been passed over for a promotion, which likely means that repeat examinees have lower average ability in domains relevant to the promotion exams. Given the high likelihood of a between-groups

ability selection effect, the results for differential item functioning in Analysis 4 will offer a clearer indication of differences in item-level performance between first-time and repeat examinees.

Item-total correlations. The differences between item-total correlations for first-time and repeat examinees are summarized in Table 44. Across all items on all tests, there was no average difference between groups' correlations. We observed small, non-significant differences for PFE items (-.02) and SKT items (.01), where negative differences indicate stronger ability discrimination for first-time examinees.

Table 44. Linear Summary Model of Differences between Corrected Item-Total Correlations for First-Time Examinees and Repeat Examinees

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	SE	p	95% CILL	95% CIUL
Model 1a: Overall Difference Effect										
Intercept	83	8,260	570,561	240,711	329,850	0.00	0.01	.61	-0.01	0.01
Model 2a: Difference Effect by Test Type										
PFE	20	1,998	478,559	205,138	273,421	-0.02	0.01	.12	-0.04	0.00
SKT	63	6,262	92,002	35,573	56,429	0.01	0.01	.14	0.00	0.02
Model 3a: Difference Effect by PFE and SKT Test										
00035	8	800	222,491	126,956	95,535	-0.03	0.02	.07	-0.06	0.00
00036	7	698	173,263	62,902	110,361	-0.01	0.02	.66	-0.04	0.02
00037	5	500	82,805	15,280	67,525	-0.01	0.02	.52	-0.05	0.03
1C151	5	500	3,703	1,790	1,913	0.00	0.02	.98	-0.04	0.04
1C171	2	199	1,153	404	749	0.02	0.03	.63	-0.05	0.08
1P051	5	499	2,051	945	1,106	-0.01	0.02	.59	-0.05	0.03
1P071	2	187	1,046	339	707	0.02	0.03	.49	-0.04	0.08
1W051	5	500	2,129	1,156	973	-0.01	0.02	.57	-0.05	0.03
1W071	4	388	2,202	493	1,709	0.02	0.02	.29	-0.02	0.07
2S051	5	500	5,665	2,614	3,051	-0.01	0.02	.77	-0.04	0.03
2S071	3	298	5,154	1,140	4,014	0.06	0.03	.01	0.01	0.11
2W151	5	500	5,491	2,651	2,840	-0.01	0.02	.49	-0.05	0.02
2W171	3	300	4,778	1,050	3,728	0.07	0.03	.00	0.02	0.12
3E751	5	500	3,087	1,500	1,587	-0.01	0.02	.67	-0.05	0.03
3E771	2	195	1,212	415	797	0.04	0.03	.17	-0.02	0.10
3P051	5	500	26,951	13,871	13,080	-0.02	0.02	.28	-0.06	0.02
3P071	4	397	18,311	3,322	14,989	0.09	0.02	.00	0.05	0.14
4N051	5	500	5,414	2,958	2,456	-0.02	0.02	.20	-0.06	0.01
4N071	3	299	3,655	925	2,730	0.01	0.03	.82	-0.04	0.05

Note. Term = Label for effect tested in the regression model; k_{Admin} = Number of test administrations that contributed to the effect; k_{Item} = Number of items that contributed to the effect; N = Total size of examinee sample; $n_{Firsttime}$ = Size of first-time examinee sample; n_{Repeat} = Size of repeat examinee sample; b = Regression coefficient indicating the average effect at a given level of aggregation; SE = Standard error of b; p = p value of for the significance test for b; 95% CILL = Lower bound of the 95% confidence interval for b; 95% CIUL = Upper bound of the 95% confidence interval for b. A positive b coefficient indicates that the average item statistic for repeat examinees was larger than for first-time examinees.

At the level of specific AFSs, only three tests had statistically significant average differences in item-total correlations. There were positive average differences in discrimination for items on the 2W171 (.07), 3P071 (.09), and 2S071 (.06) tests, the latter two of which also had significant differences in item difficulty. Positive differences indicate that, after correcting for differences in variance between groups, items for these three AFSs were more effective at discriminating high- from low-scoring examinees when the examinees had previously taken the test.

All average differences between new and exposed items were very small (≤|.02|) and did not aid in interpreting the aggregate effects (see Appendix B, Table B2).

Item-AFQT correlations. Table 45 shows that the average differences between item-AFQT correlations for first-time and repeat examinees were very small. The overall difference across all items and tests was -.02, which generalized to PFE and SKT items when they were aggregated separately. After correcting for differences in variance between the groups, the items used in promotion exams were slightly stronger discriminators of AFQT scores for first-time examinees than for repeat examinees. These differences were statistically significant, but the effects were negligible in average magnitude.

Across AFSs, the average difference in item-AFQT correlations ranged from -.07 to .03. The largest significant differences were observed for 1C171 (-.07), 1P071 (-.06), 1P051 (-.04), and 4N051 (-.04); all other effects were ≤|.03| and are reported in Table 45.

Our analyses of item-exposure effects (see Appendix B, Table B3) revealed only one difference between new and exposed items that was both statistically significant and of a non-trivial magnitude: Compared to the average -.04 difference in item-AFQT correlations for new items on the 3E771 test, the average difference was .04 (.08 higher) for exposed items.

3.3.3 Discussion

In general, the differences between item statistics for first-time examinees and repeat examinees were very small. There were only a handful of cases in which the magnitudes of mean differences between the groups' item statistics were large enough to invite scrutiny. For example, we found mean absolute differences in *p*-values as large as .17 and mean absolute differences in corrected item-total correlations as large as .09, both of which occurred for the 3P071 test. These large effects were exceptions rather than the norm, and most tests did not exhibit problematic patterns of differences.

Our statistical corrections to item-total and item-AFQT correlations ruled out key artifactual explanations for these effects. Specifically, the small or null differences between item-ability correlations from first-time and repeat examinees cannot be attributed to differences in variance between groups. We observed similar effects with and without statistical corrections, which suggests that differences in subgroup variance did not have a meaningful impact on differences between the groups' item statistics. This similarity between observed and corrected differences in correlations also supports the invariance of our conclusions under different analysis choices. Summaries of differences between the groups' item statistics treated with fewer corrections are available as a supplement to this report.

Table 45. Linear Summary Model of Differences between Item-AFQT Correlations for First-Time Examinees and Repeat Examinees

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	SE	p	95% CILL	95% CIUL
Model 1a: Overall Difference Effect										
Intercept	83	8,260	570,561	240,711	329,850	-0.02	0.00	.00	-0.03	-0.01
Model 2a: Difference Effect by Test Type										
PFE	20	1,998	478,559	205,138	273,421	-0.02	0.01	.00	-0.03	-0.01
SKT	63	6,262	92,002	35,573	56,429	-0.02	0.00	.00	-0.03	-0.01
Model 3a: Difference Effect by PFE and SKT Test										
00035	8	800	222,491	126,956	95,535	-0.03	0.01	.00	-0.04	-0.02
00036	7	698	173,263	62,902	110,361	-0.01	0.01	.10	-0.02	0.00
00037	5	500	82,805	15,280	67,525	-0.02	0.01	.01	-0.03	0.00
1C151	5	500	3,703	1,790	1,913	-0.02	0.01	.03	-0.04	0.00
1C171	2	199	1,153	404	749	-0.07	0.02	.00	-0.10	-0.04
1P051	5	499	2,051	945	1,106	-0.04	0.01	.00	-0.05	-0.02
1P071	2	187	1,046	339	707	-0.06	0.02	.00	-0.10	-0.03
1W051	5	500	2,129	1,156	973	-0.02	0.01	.03	-0.04	0.00
1W071	4	388	2,202	493	1,709	0.03	0.01	.01	0.01	0.06
2S051	5	500	5,665	2,614	3,051	-0.03	0.01	.00	-0.04	-0.01
2S071	3	298	5,154	1,140	4,014	0.01	0.01	.44	-0.01	0.03
2W151	5	500	5,491	2,651	2,840	-0.03	0.01	.00	-0.05	-0.02
2W171	3	300	4,778	1,050	3,728	0.02	0.01	.06	0.00	0.04
3E751	5	500	3,087	1,500	1,587	-0.03	0.01	.00	-0.05	-0.01
3E771	2	195	1,212	415	797	0.02	0.02	.33	-0.02	0.05
3P051	5	500	26,951	13,871	13,080	-0.03	0.01	.00	-0.04	-0.01
3P071	4	397	18,311	3,322	14,989	0.01	0.01	.28	-0.01	0.03
4N051	5	500	5,414	2,958	2,456	-0.04	0.01	.00	-0.05	-0.02
4N071	3	299	3,655	925	2,730	-0.02	0.01	.14	-0.04	0.01

Note. Term = Label for effect tested in the regression model; k_{Admin} = Number of test administrations that contributed to the effect; k_{Item} = Number of items that contributed to the effect; N = Total size of examinee sample; $n_{Firsttime}$ = Size of first-time examinee sample; n_{Repeat} = Size of repeat examinee sample; b = Regression coefficient indicating the average effect at a given level of aggregation; SE = Standard error of b; $p = p$ value of for the significance test for b; 95% CILL = Lower bound of the 95% confidence interval for b; 95% CIUL = Upper bound of the 95% confidence interval for b. A positive b coefficient indicates that the average item statistic for repeat examinees was larger than for first-time examinees.

The differences we reported between items' *p*-values posed interpretative difficulties, as simple proportion-correct statistics do not control for differences in ability between first-time and repeat examinees. These examinee groups are likely to differ in substantively important ways because, by definition, repeat examinees have previously been passed over for promotion. This type of selection artifact decreases our confidence in the assumption that the groups' *p*-values convey directly comparable information about item functioning. We address this concern in our next set

of analyses, where we evaluate differential item function (DIF) in PFE and SKT items between first-time and repeat examinees, controlling for AFQT scores.

3.4 Analysis 4: Differential Item Functioning (DIF) for First-Time and Repeat Examinees

3.4.1 Method

Sample. The sample used in this study was the same as the sample used in Analysis 3. Please see the Analysis 3 Method section for details.

DIF analyses. We used logistic regression to test for DIF between first-time and repeat examinees on each item used in each test administration. Specifically, we regressed item scores on mean-centered AFQT scores, an examinee-status dummy variable (coded so that "1" identified repeat examinees), and the interaction between AFQT scores and the dummy variable. This type of model produced a coefficient for the dummy variable that controlled for ability level and represented the effect of repeat-tester status for examinees of average ability. The natural logarithm of the dummy variable's coefficient gives the odds of repeat examinees with average ability getting the item correct relative to a first-time examinee of average ability (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990).

DIF analyses can be carried out and interpreted only when there is enough variation in both item scores and subgroup membership to estimate their association. DIF cannot be evaluated if too large a proportion of the sample belongs to just one group or if too large a proportion of examinees gets the item correct (or incorrect), as small amounts of variation preclude covariation between group membership and item scores. DIF analyses are also not meaningful for very small samples in which there are too few examinees to populate a multivariate distribution of group membership, item scores, and ability. We used three criteria to screen samples for inclusion in our summary models:

- Item difficulty (*p*-value) between .10 and .90 (inclusive)
- Proportion of first-time examinees between .10 and .90 (inclusive)
- Sample size of at least 30

These three rules focused our analyses on settings in which conducting DIF analyses was reasonable. They were effective at screening out DIF estimates that were implausibly large due to insufficient variance of one or more variables.

Mixed-effects models. We used the same method for summarizing DIF effects across items and administrations that we used for summarizing differences between item statistics in Analysis 3. We used the same set of mixed-effects linear regression models to average DIF results at different levels of aggregation and test the effect of item exposure on DIF.[10] We aggregated DIF effects in the log-odds metric in our summary models and then converted our results to the odds ratio metric for ease of interpretation.

[10] For the specifications of these models and our weighting procedures, please refer to the Method section for Analysis 3

3.4.2 Results

Table 46 shows the results of our linear summary models for DIF effects. Although some average odds ratios differed significantly from zero, all effects were very small in magnitude and did not indicate practically significant degrees of DIF. For reference, even the largest effect (1.13 for the 1P071 SKT) is equivalent to a correlation between group membership and item scores of only about .03. Based on such small effects, repeat examinees are unlikely to give correct responses at meaningfully higher rates than first-time examinees who are similar in ability.

Table 46. Linear Summary Model of Differential Item Functioning (DIF) Odds Ratios Comparing First-Time Examinees and Repeat Examinees

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	SE	p	95% CILL	95% CIUL
Model 1a: Overall DIF Effect										
Intercept	66	6,256	332,087	153,239	178,848	1.03	0.01	.01	1.01	1.05
Model 2a: DIF Effect by Test Type										
PFE	12	1,122	273,362	124,887	148,475	1.04	0.02	.06	1.00	1.09
SKT	54	5,134	58,725	28,352	30,373	1.02	0.01	.03	1.00	1.05
Model 3a: DIF Effect by PFE and SKT Test										
00035	6	558	184,036	89,620	94,416	1.04	0.03	.15	0.99	1.10
00036	3	278	55,955	22,074	33,881	1.09	0.04	.04	1.01	1.18
00037	3	286	33,371	13,193	20,178	0.99	0.04	.90	0.92	1.08
1C151	5	490	3,644	1,740	1,904	0.97	0.03	.45	0.91	1.04
1C171	1	94	178	127	51	1.02	0.09	.81	0.86	1.22
1P051	5	468	1,986	893	1,093	1.01	0.04	.81	0.94	1.08
1P071	1	97	187	122	65	1.13	0.10	.17	0.95	1.34
1W051	5	482	2,082	1,117	965	0.98	0.03	.50	0.91	1.05
1W071	3	289	629	178	451	1.04	0.05	.39	0.95	1.15
2S051	5	477	5,477	2,464	3,013	1.11	0.04	.00	1.04	1.19
2S071	1	92	585	358	227	1.04	0.08	.60	0.90	1.21
2W151	5	489	5,320	2,511	2,809	0.98	0.03	.47	0.92	1.04
2W171	2	183	1,033	368	665	1.11	0.06	.07	0.99	1.23
3E751	5	484	3,005	1,426	1,579	1.08	0.04	.03	1.01	1.15
3E771	2	174	361	136	225	0.96	0.06	.55	0.85	1.09
3P051	5	474	25,969	13,038	12,931	1.02	0.03	.56	0.96	1.08
3P071	2	176	2,248	782	1,466	0.96	0.05	.44	0.87	1.06
4N051	5	478	5,280	2,832	2,448	1.04	0.03	.20	0.98	1.11
4N071	2	187	741	260	481	1.01	0.06	.81	0.91	1.13

Note. Term = Label for effect tested in the regression model; k_{Admin} = Number of test administrations that contributed to the effect; k_{Item} = Number of items that contributed to the effect; N = Total size of examinee sample; $n_{Firsttime}$ = Size of first-time examinee sample; n_{Repeat} = Size of repeat examinee sample; b = Regression coefficient indicating the average effect at a given level of aggregation; SE = Standard error of b; p = p value of for the significance test for b; 95% CILL = Lower bound of the 95% confidence interval for b; 95% CIUL = Upper bound of the 95% confidence interval for b. A b coefficient larger than 1 indicates that, on average, repeat examinees were more likely to get items correct than were first-time examinees with similar AFQT scores.

As with our item-statistic comparisons, we explored whether items' exposure status had a moderating effect on DIF. None of the significant moderator effects revealed large differences between new and exposed items, supporting the conclusion that DIF between first-time and repeat examinees is not widespread within any test or item type (see Appendix B, Table B4).

3.4.3 Discussion

We found little evidence of DIF across promotion-exam items, and all tests' items had mean odds ratios that differed trivially from the 1.00 benchmark that indicates no effect. Our results show that first-time and repeat examinees with similar levels of ability have essentially equal odds of answering PFE and SKT items correctly.

3.5 Analysis 5: Effects of Tenure on Mean Test-Score Differences by Sex, Race, and Ethnicity

3.5.1 Method

Sample. The sample used in this study was the same as the sample used in Analysis 3. Please see the Analysis 3 Method section for details.

In addition to the data used in previous analyses, we used demographic information from Promotion Board records to determine examinees' sex, race, and Hispanic or non-Hispanic ethnicity. In our analyses of subgroup mean-difference contrasts within specific test administrations, we excluded subgroups with fewer than 20 persons.

Mean-difference analyses. We tested whether tenure had a mediating effect on the relations between scores on PFE/SKT tests and three demographic factors: sex, race, and ethnicity. We examined these effects in two different but complementary sets of analyses. The first set of analyses comprised standardized mean difference effect sizes (*d* values) computed before and after statistically controlling for TIS and TIG. The second set of analyses comprised path models that decomposed the total effect of demographic group membership on test scores into the direct effect of group membership and the indirect effect of group membership through tenure. The indirect effects from these path models reveal whether tenure explains a significant portion of the group-membership effects and help to interpret the differences between observed and statistically adjusted *d* values. We meta-analyzed the *d* values and path coefficients from each set of analyses using Schmidt and Hunter's (2015) random-effects method (as implemented in the *psychmeta* package for R) with effective sample-size weights (for information about these weights, please refer to the Method section for Analysis 3).

3.5.2 Results

The meta-analyses of *d* values are the focal point of our results. We indicate which indirect effects were statistically significant in our presentations of *d* values. We also report the meta-analyses of path coefficients in the appendix (see Appendix B, Table B5 – B16).

Across all analyses, controlling for TIG had miniscule impacts on magnitudes of subgroup mean differences. All but one of the differences between observed and adjusted *d* values were ≤ 0.02 in absolute magnitude, which is too small to be practically significant. The largest change in *d* values after controlling for TIG was -0.05 for the 2W151 SKT test's contrast between Hispanic

and non-Hispanic Airmen (a change from -0.13 to -0.08). Overall, the homogeneity of small-magnitude TIG effects suggests that TIG does not have a meaningful impact on the associations between demographic characteristics and test scores. The results for TIS also tended to be small, but the differences were more variable and will therefore be the focus of our attention.

Please note that our tables of results include some differences in *d* values with relatively large absolute-value point estimates (e.g., -0.22 for Black-White differences on the 3E771 test), but these differences tended to come from meta-analyses with small cumulative sample sizes and corresponded to indirect effects that did not differ significantly from zero. Although large in magnitude, such point-estimates of differences are not necessarily stable and should be interpreted with caution.

Sex differences. The *d* values for sex differences in Table 47 indicate that controlling for TIS had sizable effects for some tests. However, all the significant effects were positive, indicating that controlling for TIS resulted in adjusted *d* values that were lower in raw value than their observed counterparts (i.e., negative effects were made stronger by the adjustment, or positive effects were made weaker). These differences suggest suppression effects for SKT overall (0.01), 2W171 (0.14), and 3P071 (0.05) because the differences favoring men were larger after controlling for TIS. A small difference favoring women on the 2S071 test become smaller by 0.07 after controlling for TIS.

Minority-White differences. In terms of minority-White mean differences, controlling for TIS had greater impact on Black-White differences than on any other groupwise contrast (see Table 48). Controlling for TIS yielded small reductions in *d* values across all tests (-0.03), as well as PFE tests (-0.03) and SKT tests (-0.02) when they were aggregated separately. The impact of TIS was more variable across PFE and SKT tests, with 00037 showing the largest difference in *d* values (-0.10) out of all significant indirect main effects; this effect is also notable because it represented a drop from a small difference (d = -0.12) to a near-zero difference (d = -0.02) for a highly powered sample. The 3P071 test also had a sizable reduction in standardized mean differences after controlling for TIS (-0.08), a change from an observed *d* of -0.50 to an adjusted estimate of -0.42. Controlling for the effect of TIS had smaller impact on *d* values for three other tests, namely 1P071 (-0.05), 2W171 (-0.04), and 4N071 (-0.03).

The changes in *d* values were small for all other minority-White contrasts with significant indirect effects. Only the TIS indirect effect for the 00037 PFE test was statistically significant for Asian-White differences (see Table 49), but the difference in *d* values associated with this effect was trivial in magnitude (-0.01). There were no significant indirect effects for Pacific Islander-White differences (see Table 50) and all the *d*-value differences for these comparisons were between -0.01 and 0.01. Controlling for TIS in American Indian-White differences had effects that were limited to PFE tests (see Table 51) and the average differences between observed and adjusted *d* values were small for both the 00035 (-0.00) and 00037 (-0.03) tests.

Hispanic-non-Hispanic d values. For Hispanic-non-Hispanic contrasts (see Table 52), the only significant indirect effect for TIS occurred on the 2W151 test and the difference between that test's observed and adjusted *d* values was -0.07 (a change in magnitude from -0.13 to -0.06). Note that 2W151 was also the only test that produced a change in *d* values larger than 0.02 in absolute magnitude after controlling for TIG (-0.05; a change from -0.13 to -0.08). This suggests that, although the changes in mean-difference magnitude were small, tenure effects may be more pervasive in the 2W151 AFSC than in others.

Table 47. Meta-Analyses of Female-Male Standardized Mean Differences in Observed PFE/SKT Test Scores, Test Scores Controlling for Time in Grade (TIG), and Test Scores Controlling for Time in Grade (TIS)

					Observed *d* values		TIG-Controlled *d* values			TIS-Controlled *d* values		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	***Diff.***	**95% CI**	***Mean***	***Diff.***	**95% CI**
Overall	99	525,201	106,233	418,968	−0.06	(−0.09, −0.03)	−0.06	0.00	(−0.08, −0.03)	−0.06	0.00	(−0.09, −0.03)
PFE	23	458,328	90,735	367,593	−0.03	(−0.07, 0.00)	−0.03	0.00	(−0.07, 0.00)	−0.03	−0.00	(−0.07, 0.00)
00035	9	207,965	40,606	167,359	−0.02	(−0.07, 0.04)	−0.02	0.00	(−0.07, 0.04)	−0.02	0.00	(−0.07, 0.04)
00036	7	154,934	31,081	123,853	−0.07	(−0.11, −0.02)	−0.07	0.00	(−0.12, −0.02)	−0.06	−0.00	(−0.11, −0.02)
00037	7	95,429	19,048	76,381	−0.01	(−0.13, 0.10)	−0.01	−0.00	(−0.13, 0.10)	−0.01	−0.00	(−0.13, 0.11)
SKT	76	66,873	15,498	51,375	−0.24	(−0.29, −0.19)	−0.24	**−0.00**	(−0.29, −0.19)	−0.25	**0.01**	(−0.30, −0.20)
1C151	6	3,854	588	3,266	−0.13	(−0.20, −0.07)	−0.14	0.00	(−0.20, −0.07)	−0.14	0.01	(−0.21, −0.08)
1C171	5	1,057	209	848	−0.00	(−0.25, 0.24)	−0.02	0.01	(−0.25, 0.22)	−0.05	0.05	(−0.26, 0.16)
1P051	6	2,044	344	1,700	−0.12	(−0.24, 0.01)	−0.11	−0.00	(−0.23, 0.01)	−0.11	−0.00	(−0.23, 0.00)
1P071	5	949	202	747	−0.26	(−0.42, −0.10)	−0.25	−0.01	(−0.42, −0.08)	−0.31	0.05	(−0.52, −0.10)
1W051	6	2,079	518	1,561	−0.42	(−0.59, −0.24)	−0.41	−0.00	(−0.58, −0.25)	−0.42	0.00	(−0.59, −0.25)
1W071	5	970	149	821	−0.40	(−0.57, −0.22)	−0.40	0.00	(−0.58, −0.22)	−0.51	0.11	(−0.71, −0.31)
2S051	6	5,750	2,024	3,726	0.04	(−0.04, 0.11)	0.04	−0.00	(−0.04, 0.11)	0.04	−0.00	(−0.03, 0.11)
2S071	5	2,802	1,265	1,537	0.11	(0.02, 0.20)	0.11	0.00	(0.02, 0.20)	0.04	**0.07**	(−0.05, 0.14)
2W151	6	5,598	761	4,837	−0.52	(−0.61, −0.43)	−0.52	0.00	(−0.62, −0.41)	−0.53	0.01	(−0.63, −0.42)
2W171	4	2,121	146	1,975	−0.45	(−0.76, −0.15)	−0.45	**−0.00**	(−0.75, −0.14)	−0.60	**0.14**	(−0.88, −0.31)
3P051	6	26,460	4,784	21,676	−0.37	(−0.44, −0.29)	−0.36	−0.01	(−0.44, −0.28)	−0.36	−0.00	(−0.44, −0.28)
3P071	5	5,820	555	5,265	−0.50	(−0.62, −0.37)	−0.49	−0.00	(−0.61, −0.38)	−0.55	**0.05**	(−0.67, −0.43)
4N051	6	5,424	2,944	2,480	−0.20	(−0.30, −0.11)	−0.20	0.00	(−0.30, −0.10)	−0.20	0.00	(−0.31, −0.10)
4N071	5	1,945	1,009	936	0.05	(0.01, 0.10)	0.06	−0.00	(0.01, 0.10)	0.02	0.03	(−0.03, 0.07)

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average *d* value; Diff. = Difference between observed mean *d* value and statistically adjusted mean *d* value, 95% CI = 95% confidence interval around the meta-analytic mean. Bolded "Diff." values indicate a statistically significant indirect effect. Negative *d* values indicate that the referent group's mean was higher than the focal group's mean.

Table 48. Meta-Analyses of Black-White Standardized Mean Differences in Observed PFE/SKT Test Scores, Test Scores Controlling for Time in Grade (TIG), and Test Scores Controlling for Time in Grade (TIS)

					Observed *d* values		TIG-Controlled *d* values			TIS-Controlled *d* values		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	***Diff.***	**95% CI**	***Mean***	***Diff.***	**95% CI**
Overall	106	372,092	75,024	297,068	−0.23	(−0.25, −0.21)	−0.23	**0.00**	(−0.25, −0.21)	−0.20	**−0.03**	(−0.23, −0.17)
PFE	23	325,414	64,010	261,404	−0.21	(−0.24, −0.17)	−0.21	**0.00**	(−0.25, −0.17)	−0.18	**−0.03**	(−0.22, −0.13)
00035	9	139,922	26,776	113,146	−0.23	(−0.28, −0.17)	−0.22	−0.00	(−0.28, −0.17)	−0.22	−0.00	(−0.28, −0.16)
00036	7	116,630	22,706	93,924	−0.24	(−0.27, −0.20)	−0.25	**0.01**	(−0.28, −0.21)	−0.22	−0.02	(−0.25, −0.19)
00037	7	68,862	14,528	54,334	−0.12	(−0.24, −0.01)	−0.12	−0.00	(−0.23, −0.01)	−0.02	**−0.10**	(−0.12, 0.07)
SKT	83	46,678	11,014	35,664	−0.38	(−0.42, −0.34)	−0.38	−0.01	(−0.42, −0.34)	−0.36	**−0.02**	(−0.40, −0.32)
1C151	6	2,581	330	2,251	−0.23	(−0.35, −0.12)	−0.22	−0.01	(−0.32, −0.11)	−0.22	−0.02	(−0.33, −0.11)
1C171	5	844	154	690	−0.29	(−0.45, −0.13)	−0.29	−0.01	(−0.42, −0.15)	−0.23	−0.06	(−0.36, −0.10)
1P051	6	1,391	274	1,117	−0.32	(−0.41, −0.23)	−0.32	0.00	(−0.42, −0.23)	−0.32	0.00	(−0.41, −0.23)
1P071	5	694	207	487	−0.37	(−0.59, −0.15)	−0.37	−0.00	(−0.60, −0.14)	−0.32	**−0.05**	(−0.51, −0.12)
1W051	4	1,302	149	1,153	−0.42	(−0.60, −0.25)	−0.41	−0.01	(−0.60, −0.21)	−0.40	−0.02	(−0.57, −0.23)
1W071	3	523	67	456	−0.56	(−0.70, −0.41)	−0.55	−0.00	(−0.63, −0.48)	−0.40	−0.15	(−0.66, −0.15)
2S051	6	3,617	1,478	2,139	−0.11	(−0.20, −0.01)	−0.11	0.01	(−0.22, −0.00)	−0.11	0.00	(−0.21, −0.01)
2S071	5	1,757	979	778	−0.00	(−0.04, 0.03)	−0.00	−0.00	(−0.03, 0.03)	0.01	−0.02	(−0.07, 0.10)
2W151	6	3,743	704	3,039	−0.43	(−0.55, −0.30)	−0.41	−0.02	(−0.52, −0.30)	−0.40	−0.02	(−0.51, −0.29)
2W171	5	1,894	431	1,463	−0.38	(−0.55, −0.22)	−0.38	0.00	(−0.55, −0.22)	−0.34	**−0.04**	(−0.48, −0.20)
3E751	6	2,025	319	1,706	−0.55	(−0.66, −0.43)	−0.53	−0.01	(−0.66, −0.41)	−0.53	−0.01	(−0.65, −0.42)
3E771	4	616	106	510	−0.61	(−0.80, −0.43)	−0.63	0.01	(−0.80, −0.45)	−0.40	−0.22	(−0.64, −0.15)
3P051	6	16,635	3,724	12,911	−0.50	(−0.56, −0.45)	−0.49	−0.01	(−0.54, −0.44)	−0.49	−0.01	(−0.55, −0.43)
3P071	5	4,202	847	3,355	−0.50	(−0.66, −0.34)	−0.50	−0.00	(−0.66, −0.33)	−0.42	**−0.08**	(−0.61, −0.23)
4N051	6	3,617	840	2,777	−0.38	(−0.51, −0.25)	−0.37	−0.01	(−0.50, −0.24)	−0.37	−0.01	(−0.49, −0.24)
4N071	5	1,237	405	832	−0.20	(−0.39, −0.02)	−0.21	0.00	(−0.39, −0.02)	−0.17	**−0.03**	(−0.37, 0.02)

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average *d* value; Diff. = Difference between observed mean *d* value and statistically adjusted mean *d* value, 95% CI = 95% confidence interval around the meta-analytic mean. Bolded "Diff." values indicate a statistically significant indirect effect. Negative *d* values indicate that the referent group's mean was higher than the focal group's mean.

Table 49. Meta-Analyses of Asian-White Standardized Mean Differences in Observed PFE/SKT Test Scores, Test Scores Controlling for Time in Grade (TIG), and Test Scores Controlling for Time in Grade (TIS)

					Observed *d* values		**TIG-Controlled *d* values**			**TIS-Controlled *d* values**		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	***Diff.***	**95% CI**	***Mean***	***Diff.***	**95% CI**
Overall	52	295,487	12,670	282,817	−0.13	(−0.16, −0.10)	−0.13	**0.00**	(−0.16, −0.10)	−0.13	−0.00	(−0.16, −0.10)
PFE	23	273,010	11,606	261,404	−0.13	(−0.17, −0.09)	−0.13	**0.00**	(−0.17, −0.09)	−0.13	−0.00	(−0.17, −0.09)
00035	9	118,252	5,106	113,146	−0.08	(−0.18, 0.01)	−0.08	0.00	(−0.18, 0.01)	−0.09	0.00	(−0.18, 0.01)
00036	7	98,151	4,227	93,924	−0.16	(−0.21, −0.11)	−0.16	**0.00**	(−0.21, −0.12)	−0.16	−0.00	(−0.21, −0.11)
00037	7	56,607	2,273	54,334	−0.17	(−0.25, −0.09)	−0.17	−0.00	(−0.25, −0.09)	−0.17	**−0.01**	(−0.25, −0.09)
SKT	29	22,477	1,064	21,413	−0.12	(−0.21, −0.02)	−0.11	**−0.01**	(−0.20, −0.02)	−0.12	0.01	(−0.21, −0.03)
2S051	6	2,361	222	2,139	0.13	(−0.08, 0.34)	0.13	−0.00	(−0.07, 0.34)	0.13	−0.00	(−0.07, 0.34)
2S071	2	388	49	339	0.39	(−0.25, 1.03)	0.38	0.00	(−0.19, 0.95)	0.28	0.11	(−0.00, 0.56)
2W151	5	2,858	130	2,728	−0.09	(−0.23, 0.05)	−0.07	−0.02	(−0.21, 0.07)	−0.08	−0.02	(−0.21, 0.06)
2W171	3	1,101	77	1,024	−0.44	(−0.76, −0.13)	−0.44	−0.00	(−0.76, −0.12)	−0.49	0.04	(−0.73, −0.24)
3P051	5	12,316	319	11,997	−0.30	(−0.44, −0.17)	−0.30	−0.01	(−0.44, −0.16)	−0.30	−0.00	(−0.44, −0.16)
4N051	6	2,997	220	2,777	−0.09	(−0.18, −0.00)	−0.09	**−0.00**	(−0.18, 0.00)	−0.09	−0.00	(−0.18, 0.00)
4N071	2	456	47	409	−0.05	(−2.03, 1.94)	−0.04	−0.00	(−2.01, 1.92)	−0.11	0.07	(−2.20, 1.97)

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average *d* value; Diff. = Difference between observed mean *d* value and statistically adjusted mean *d* value, 95% CI = 95% confidence interval around the meta-analytic mean. Bolded "Diff." values indicate a statistically significant indirect effect. Negative *d* values indicate that the referent group's mean was higher than the focal group's mean.

Table 50. Meta-Analyses of Pacific Islander-White Standardized Mean Differences in Observed PFE/SKT Test Scores, Test Scores Controlling for Time in Grade (TIG), and Test Scores Controlling for Time in Grade (TIS)

					Observed *d* values		**TIG-Controlled *d* values**			**TIS-Controlled *d* values**		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	***Diff.***	**95% CI**	***Mean***	***Diff.***	**95% CI**
Overall	29	280,183	5,868	274,315	−0.23	(−0.27, −0.20)	−0.23	−0.00	(−0.27, −0.20)	−0.23	−0.01	(−0.26, −0.19)
PFE	23	267,004	5,600	261,404	−0.23	(−0.27, −0.20)	−0.23	−0.00	(−0.27, −0.20)	−0.23	−0.01	(−0.26, −0.19)
00035	9	115,370	2,224	113,146	−0.21	(−0.29, −0.13)	−0.21	−0.00	(−0.28, −0.13)	−0.21	−0.00	(−0.28, −0.13)
00036	7	96,245	2,321	93,924	−0.27	(−0.30, −0.24)	−0.27	0.00	(−0.30, −0.24)	−0.27	−0.00	(−0.30, −0.24)
00037	7	55,389	1,055	54,334	−0.20	(−0.27, −0.13)	−0.20	−0.00	(−0.26, −0.13)	−0.18	−0.01	(−0.25, −0.12)
SKT	6	13,179	268	12,911	−0.25	(−0.50, −0.00)	−0.23	−0.01	(−0.48, 0.02)	−0.24	−0.01	(−0.49, 0.01)
3P051	6	13,179	268	12,911	−0.25	(−0.50, −0.00)	−0.23	−0.01	(−0.48, 0.02)	−0.24	−0.01	(−0.49, 0.01)

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average *d* value; Diff. = Difference between observed mean *d* value and statistically adjusted mean *d* value, 95% CI = 95% confidence interval around the meta-analytic mean. Bolded "Diff." values indicate a statistically significant indirect effect. Negative *d* values indicate that the referent group's mean was higher than the focal group's mean.

Table 51. Meta-Analyses of American Indian-White Standardized Mean Differences in Observed PFE/SKT Test Scores, Test Scores Controlling for Time in Grade (TIG), and Test Scores Controlling for Time in Grade (TIS)

					Observed *d* values		**TIG-Controlled *d* values**			**TIS-Controlled *d* values**		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	***Diff.***	**95% CI**	***Mean***	***Diff.***	**95% CI**
Overall	25	270,780	2,780	268,000	−0.22	(−0.28, −0.17)	−0.23	**0.00**	(−0.28, −0.18)	−0.21	−0.01	(−0.27, −0.16)
PFE	21	259,952	2,641	257,311	−0.21	(−0.26, −0.17)	−0.22	**0.00**	(−0.27, −0.17)	−0.20	−0.01	(−0.26, −0.15)
00035	9	114,283	1,137	113,146	−0.23	(−0.29, −0.16)	−0.23	**0.00**	(−0.30, −0.16)	−0.23	**−0.00**	(−0.29, −0.16)
00036	7	94,936	1,012	93,924	−0.21	(−0.31, −0.11)	−0.22	**0.01**	(−0.32, −0.12)	−0.20	−0.01	(−0.31, −0.10)
00037	5	50,733	492	50,241	−0.19	(−0.40, 0.01)	−0.19	−0.00	(−0.40, 0.02)	−0.16	**−0.03**	(−0.39, 0.07)
SKT	4	10,828	139	10,689	−0.43	(−0.82, −0.04)	−0.41	−0.02	(−0.80, −0.02)	−0.42	−0.01	(−0.80, −0.03)
3P051	4	10,828	139	10,689	−0.43	(−0.82, −0.04)	−0.41	−0.02	(−0.80, −0.02)	−0.42	−0.01	(−0.80, −0.03)

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average *d* value; Diff. = Difference between observed mean *d* value and statistically adjusted mean *d* value, 95% CI = 95% confidence interval around the meta-analytic mean. Bolded "Diff." values indicate a statistically significant indirect effect. Negative *d* values indicate that the referent group's mean was higher than the focal group's mean.

Table 52. Meta-Analyses of Hispanic-Non-Hispanic Standardized Mean Differences in Observed PFE/SKT Test Scores, Test Scores Controlling for Time in Grade (TIG), and Test Scores Controlling for Time in Grade (TIS)

					Observed *d* values		**TIG-Controlled *d* values**			**TIS-Controlled *d* values**		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	***Diff.***	**95% CI**	***Mean***	***Diff.***	**95% CI**
Overall	72	515,904	26,080	489,824	−0.09	(−0.11, −0.08)	−0.10	0.01	(−0.12, −0.08)	−0.07	−0.02	(−0.09, −0.06)
PFE	23	458,328	23,388	434,940	−0.10	(−0.12, −0.07)	−0.10	0.01	(−0.13, −0.08)	−0.08	−0.02	(−0.09, −0.06)
00035	9	207,965	5,236	202,729	−0.07	(−0.12, −0.02)	−0.07	−0.00	(−0.11, −0.02)	−0.07	−0.00	(−0.11, −0.02)
00036	7	154,934	9,035	145,899	−0.13	(−0.15, −0.10)	−0.14	0.02	(−0.17, −0.12)	−0.10	−0.02	(−0.13, −0.08)
00037	7	95,429	9,117	86,312	−0.08	(−0.14, −0.03)	−0.08	−0.00	(−0.13, −0.03)	−0.05	−0.03	(−0.08, −0.03)
SKT	49	57,576	2,692	54,884	−0.08	(−0.12, −0.03)	−0.07	−0.01	(−0.11, −0.03)	−0.07	−0.01	(−0.11, −0.03)
1C171	3	671	69	602	0.11	(−0.14, 0.36)	0.12	−0.01	(−0.01, 0.25)	0.17	−0.05	(−0.00, 0.34)
1P071	1	208	21	187	0.30	(−0.16, 0.75)	0.30	−0.01	(−0.15, 0.75)	0.30	−0.01	(−0.15, 0.76)
2S051	6	5,750	244	5,506	0.01	(−0.24, 0.27)	0.02	−0.01	(−0.23, 0.27)	0.02	−0.00	(−0.24, 0.27)
2S071	5	2,802	344	2,458	0.09	(−0.09, 0.27)	0.09	0.00	(−0.08, 0.26)	0.07	0.02	(−0.07, 0.20)
2W151	6	5,598	165	5,433	−0.13	(−0.18, −0.07)	−0.08	**−0.05**	(−0.16, −0.00)	−0.06	**−0.07**	(−0.14, 0.02)
2W171	5	2,681	235	2,446	−0.03	(−0.12, 0.06)	−0.03	−0.00	(−0.12, 0.06)	−0.02	−0.01	(−0.12, 0.08)
3E771	1	217	20	197	−0.20	(−0.66, 0.27)	−0.17	−0.02	(−0.63, 0.29)	−0.09	−0.11	(−0.55, 0.38)
3P051	6	26,460	695	25,765	−0.15	(−0.23, −0.06)	−0.14	−0.01	(−0.22, −0.06)	−0.14	−0.01	(−0.23, −0.05)
3P071	5	5,820	522	5,298	−0.14	(−0.25, −0.04)	−0.15	0.00	(−0.26, −0.04)	−0.13	−0.01	(−0.21, −0.05)
4N051	6	5,424	137	5,287	−0.01	(−0.19, 0.17)	−0.00	−0.01	(−0.16, 0.16)	−0.01	−0.01	(−0.16, 0.15)
4N071	5	1,945	240	1,705	−0.16	(−0.29, −0.03)	−0.16	0.00	(−0.29, −0.03)	−0.16	0.00	(−0.32, −0.00)

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average *d* value; Diff. = Difference between observed mean *d* value and statistically adjusted mean *d* value, 95% CI = 95% confidence interval around the meta-analytic mean. Bolded "Diff." values indicate a statistically significant indirect effect. Negative *d* values indicate that the referent group's mean was higher than the focal group's mean.

3.5.3 Discussion

In most analyses, controlling for tenure had neither a practically nor statistically significant effect on magnitudes of subgroup mean differences. The effects of TIG on mean differences were zero or very near zero in all analyses, and TIS had non-zero effects of non-trivial magnitudes in only a handful of cases. When subgroup membership had non-zero indirect effects on test scores through tenure, the differences were such that the observed and statistically adjusted standardized mean differences both represented the same general magnitude of effect – there was no case in which a moderate or large difference was reduced to a small or zero difference. Based on these results, it is unlikely that either TIG or TIS systematically contributes to the observed mean differences in promotion test scores among demographic subgroups.

4.0 SUMMARY

The Air Force's WAPS testing program has experienced a reduction in scope as a result of recent policy decisions initiated by Air Force leadership. Analyses conducted during Tasks 1 and 2 of the current project addressed some of the key psychometric characteristics of the WAPS tests (SKT and PFE). The criterion-related validity analyses from Task 1 provided modest evidence at best for the validity of WAPS scores for predicting future job performance. Nevertheless, the psychometric characteristics of these measures seems strong. The Task 2 analyses unearthed very little evidence that would justify concerns regarding attenuated validity due to differential performance on items as a function of exposure rates. Further, few differences by demographic group or experience were identified with regard to these effects.

Future research efforts should concentrate on development of improved occupational performance criteria for use in criterion-related validity studies of the WAPS tests. Current practice regarding item retesting does not seem to lead to compromised psychometric characteristics of the WAPS measures. Even so, future research could be conducted on larger sample sizes to help dampen some of the coefficient "bounce" we observed across AFSCs or even finer cuts of the data resulting in samples uniform with regard to AFSC-by-cycle or AFSC-by-grade classifications.

5.0 REFERENCES

Bakbergenuly, I., Hoaglin, D. C., & Kulinskaya, E. (2019a). Simulation study of estimating between-study variance and overall effect in meta-analysis of odds-ratios. *ArXiv:1902.07154 [Stat]*. Retrieved from http://arxiv.org/abs/1902.07154

Bakbergenuly, I., Hoaglin, D. C., & Kulinskaya, E. (2019b). Simulation study of estimating between-study variance and overall effect in meta-analysis of standardized mean difference. *ArXiv:1903.01362 [Stat]*. Retrieved from http://arxiv.org/abs/1903.01362

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (August 25, 1978). *Uniform guidelines on employee selection procedures*. Washington, DC: Author.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2*, 1360-1383. http://doi.org/10.1214/08-AOAS191

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh*, *62*(A), 28-30.

Losey, S. (2019, February 4). *Air Force drops WAPS testing for SNCOs*. Retrieved from https://www.airforcetimes.com/news/your-air-force/2019/02/04/air-force-drops-waps-testing-for-sncos/

Maier, M. H., & Grafton, F. C. (1981). *Aptitude composites for ASVAB 8, 9, and 10*. Alexandria, VA: U.S. Army Research institute for the Behavioral and Social Sciences.

McCloy, R.A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology*, *79*(4), 493-505.

Oppler, S. H., McCloy, R. A., & Campbell, J. P. (2001). The prediction of supervisory and leadership performance. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 389-409). Mahwah, NJ: Lawrence Erlbaum Associates.

Oppler, S. H., McCloy, R. A., Peterson, N. G., Russell, T. L., & Campbell, J. P. (2001). The prediction of multiple components of entry-level performance. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 349-388). Mahwah, NJ: Erlbaum.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105–116. https://doi.org/10.1177/014662169301700201

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.

Society for Industrial and Organizational Psychology, Inc. (2018), *Principles for the validation and use of personnel selection procedures* (5th Ed.). Bowling Green, OH: Author.

Stan Development Team. (2016). *rstanarm: Bayesian applied regression modeling via Stan.* R package version 2.13.1, http://mc-stan.org/.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Welsh, J. R., Kucinkas, S. K., & Curran, L. T. (1990). *Armed Services Vocational Battery (ASVAB): Integrative review of validity studies.* Brooks AFB, TX: Air Force Human Resources Laboratory.

APPENDIX A – ADDITIONAL TASK 1 RESULTS

Exhibit A1. Descriptive Statistics for Control Variables

	Promotion EPR			Decorations			Time in Grade			Time in Service		
	n	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***	**Mean**	**SD**	***n***	**Mean**	**SD**
Overall	27,470	168.70	59.08	27,470	3.32	4.44	27,470	18.20	13.70	27,470	12.30	9.63
Grade												
E5	16,371	177.39	48.81	16,371	.65	.95	16,371	12.54	8.16	16,371	7.64	4.54
E6	7,121	174.96	49.65	7,121	5.12	3.17	7,121	28.12	17.03	7,121	16.83	9.73
E7	3,978	121.72	85.46	3,978	11.07	4.41	3,978	23.77	12.79	3,978	23.39	11.70
Cycle												
11	4,214	132.52	5.23	4,214	3.47	4.29	4,214	26.80	13.00	4,214	18.02	8.68
12	4,586	132.26	5.31	4,586	3.25	4.23	4,586	25.85	12.55	4,586	17.52	8.75
13	3,374	132.11	5.59	3,374	3.11	4.38	3,374	25.42	11.98	3,374	17.38	9.04
14	2,227	132.54	4.86	2,227	3.85	4.90	2,227	28.11	11.86	2,227	19.39	9.57
15	4,260	249.09	3.32	4,260	3.81	4.94	4,260	17.72	7.38	4,260	11.82	5.64
16	5,599	189.83	74.29	5,599	2.91	4.21	5,599	7.99	3.49	5,599	5.26	2.61
17	3,210	188.24	66.65	3,210	3.14	4.28	3,210	.00	.00	3,210	.00	.00
AFSC												
1C1X1	1,912	166.26	59.02	1,912	3.02	4.35	1,912	17.11	11.97	1,912	12.04	9.01
1P0X1	1,216	167.23	60.46	1,216	3.60	4.24	1,216	19.71	14.08	1,216	13.56	10.08
1W0X1	1,420	166.38	61.19	1,420	4.34	5.28	1,420	18.31	13.48	1,420	13.12	10.15
2S0X1	3,372	163.65	61.81	3,372	4.05	4.77	3,372	20.96	15.29	3,372	14.26	10.73
2W1X1	3,275	167.68	62.15	3,275	3.20	3.96	3,275	19.50	14.55	3,275	13.18	10.16
3E7X1	1,499	167.25	58.30	1,499	2.87	3.93	1,499	18.12	13.63	1,499	12.16	9.56
3P0X1	11,874	172.04	57.17	11,874	3.18	4.47	11,874	16.91	12.77	11,874	11.25	8.87
4N0X1	2,902	166.13	58.09	2,902	2.98	4.25	2,902	18.92	14.70	2,902	12.66	10.07

Exhibit A2. Distributions of predictor-outcome correlations across all AFSC, cycles, and grades

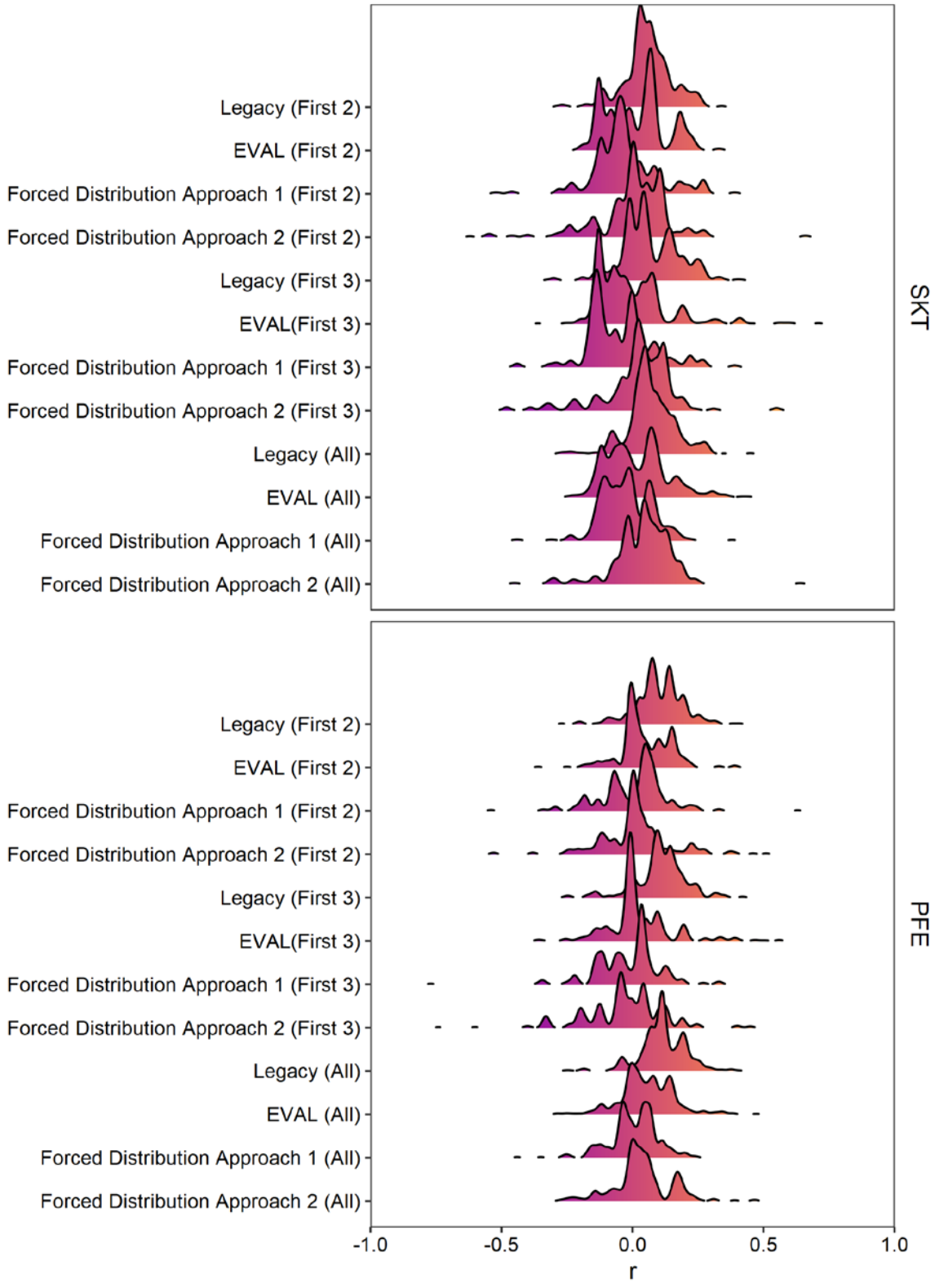


Exhibit A3. Distributions of predictor-control correlations across all AFSC, cycles, and grades.

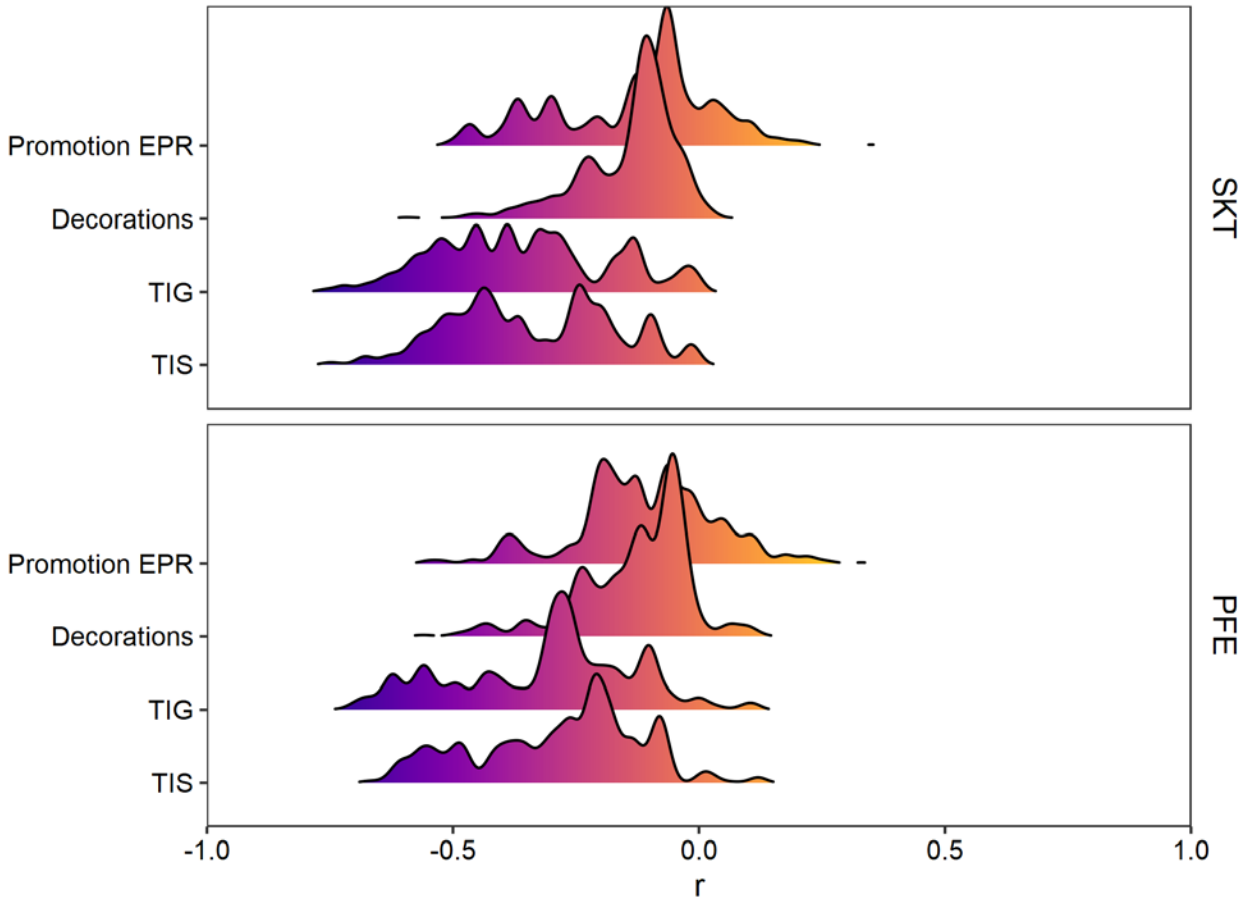


Exhibit A4. Distributions of predictor-control correlations across all AFSC, cycles, and grades, corrected for multivariate range restriction.

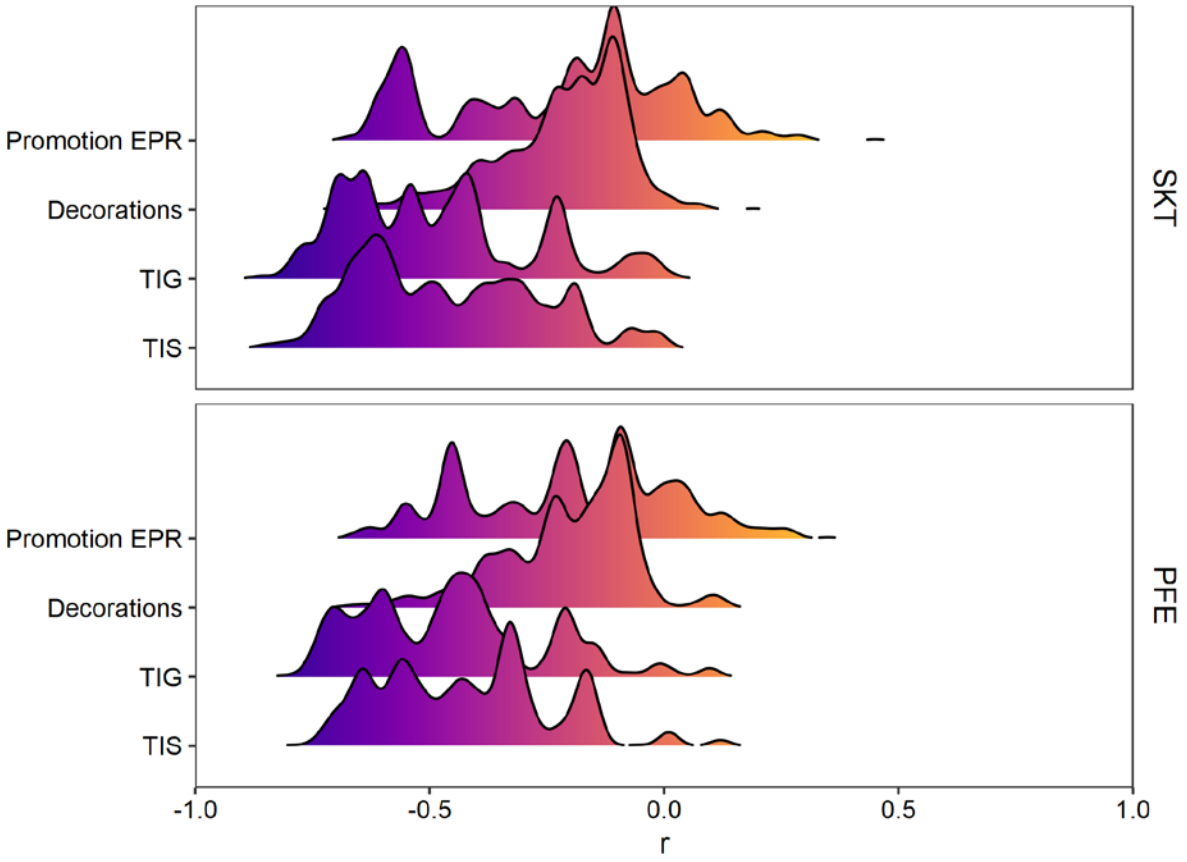


Exhibit A5. Distributions of control-outcome correlations across all AFSC, cycles, and grades.

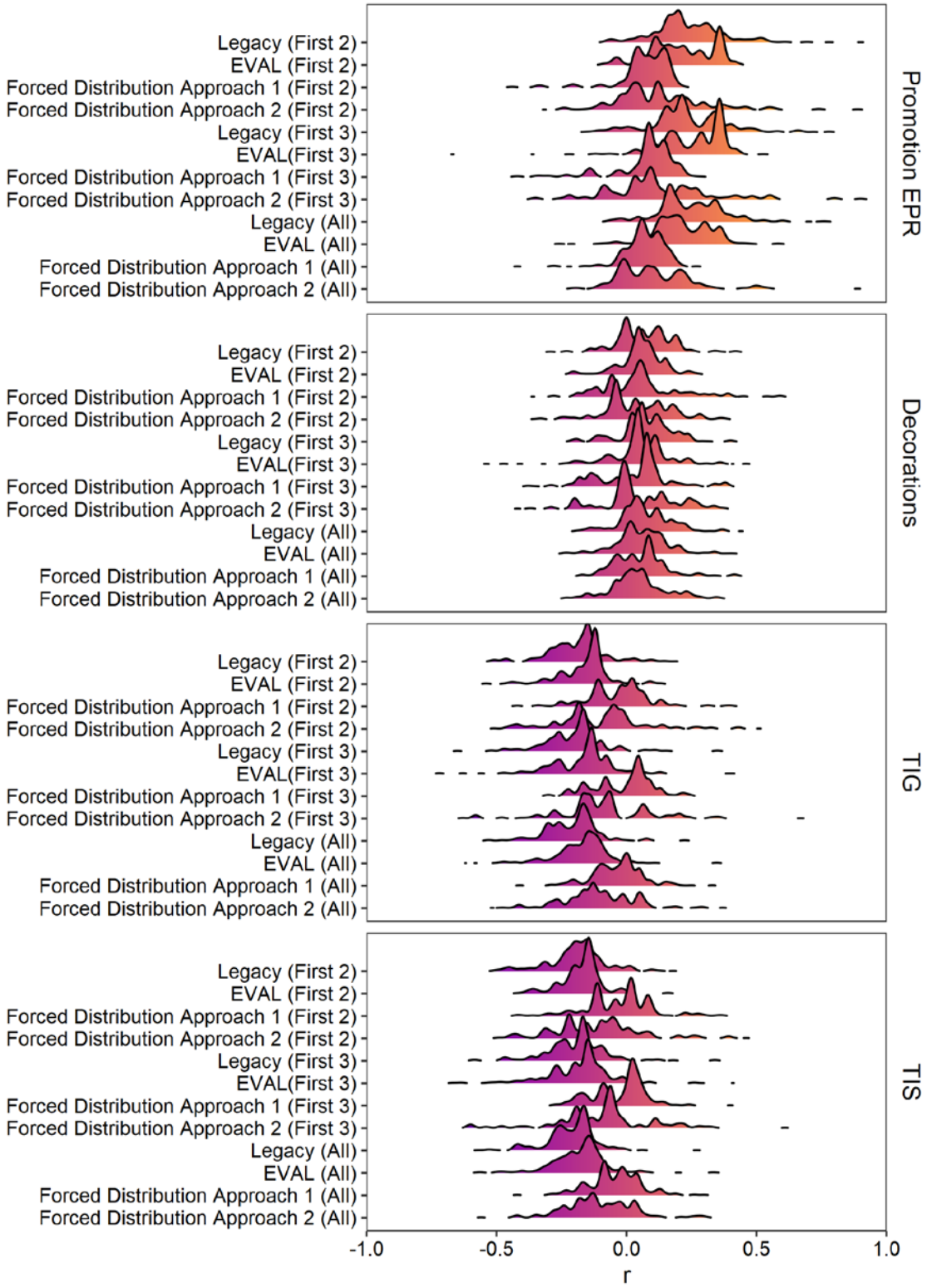


Exhibit A6. Distributions of control-outcome correlations across all AFSC, cycles, and grades, corrected for multivariate range restriction.

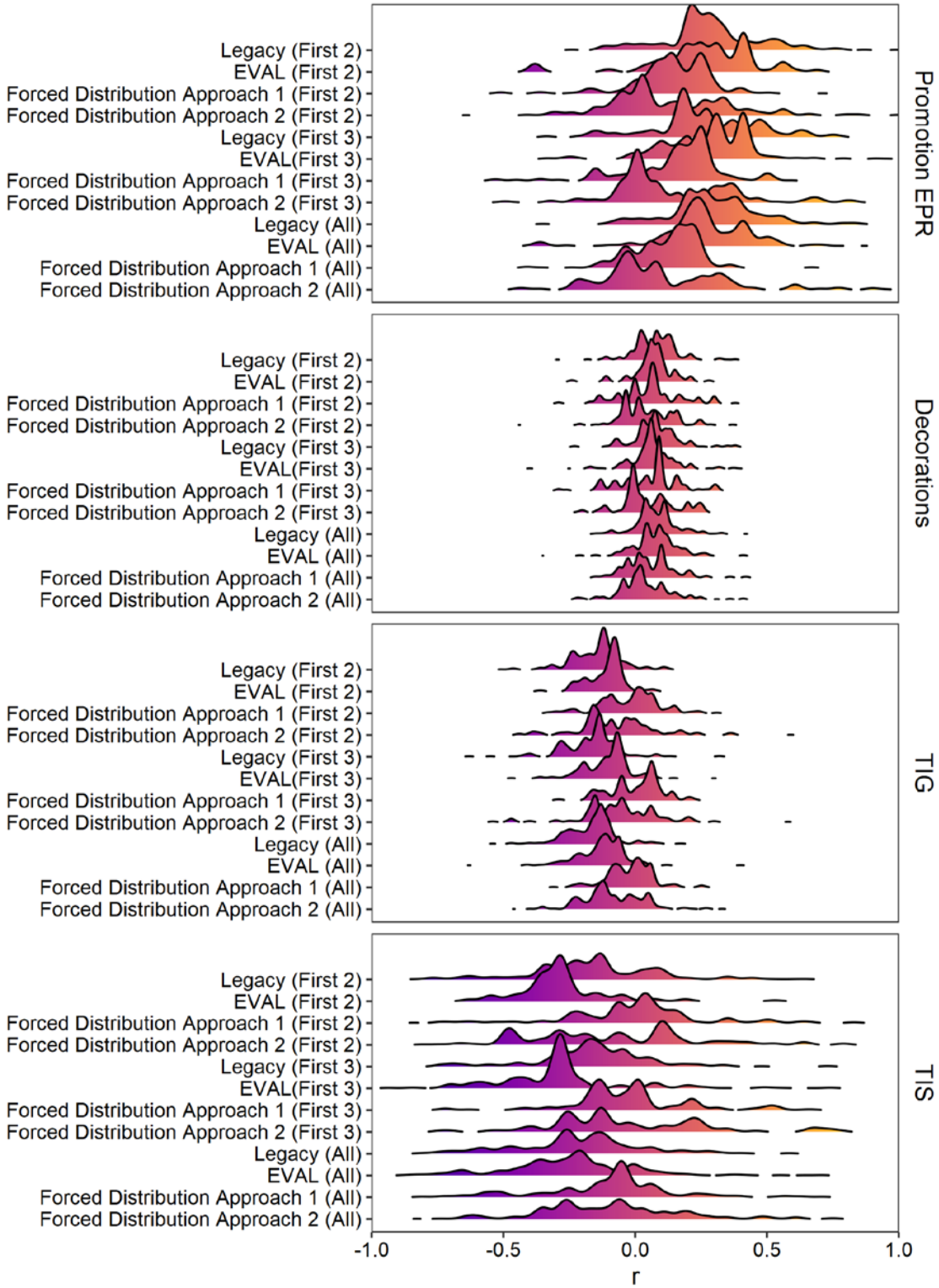


Exhibit A7. Descriptive Statistics for ASVAB Variables

	AFQT			MAGE M			MAGE A			MAGE G			MAGE E		
	n	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**
Overall	27,045	63.78	16.59	27,758	60.04	21.27	27,758	65.97	15.98	27,758	63.51	17.42	27,758	64.93	18.60
Grade															
E5	16,174	66.03	15.86	16,433	61.02	20.59	16,433	67.72	15.13	16,433	65.09	17.04	16,433	66.79	18.06
E6	7,006	60.70	17.15	7,317	58.44	21.91	7,317	63.35	16.54	7,317	61.06	17.85	7,317	62.26	19.15
E7	3,865	59.98	16.98	4,008	58.96	22.53	4,008	63.59	17.31	4,008	61.49	17.42	4,008	62.20	18.80
Cycle															
11	4,088	60.92	16.81	4,208	59.17	21.81	4,208	63.75	16.61	4,208	61.36	17.68	4,208	62.92	18.83
12	4,468	61.80	17.04	4,618	59.04	21.47	4,618	64.26	16.40	4,618	62.14	17.64	4,618	63.15	18.77
13	3,308	63.25	17.27	3,416	60.10	21.63	3,416	65.13	16.73	3,416	63.34	17.98	3,416	64.54	19.17
14	2,238	63.41	16.90	2,301	59.23	21.68	2,301	65.81	16.32	2,301	63.36	17.68	2,301	63.94	19.17
15	4,213	64.78	16.59	4,333	61.11	21.33	4,333	66.90	15.75	4,333	64.24	17.51	4,333	65.92	18.54
16	5,558	66.75	15.92	5,648	61.59	20.86	5,648	68.57	15.17	5,648	65.81	16.95	5,648	67.51	18.08
17	3,172	64.58	14.79	3,234	59.00	19.92	3,234	66.54	14.34	3,234	63.51	16.07	3,234	65.42	17.30
AFSC															
1C1X1	1,917	78.37	12.87	1,929	76.07	15.94	1,929	78.15	13.66	1,929	78.61	12.46	1,929	79.47	13.80
1P0X1	1,170	58.83	15.69	1,230	57.55	18.46	1,230	61.60	15.93	1,230	59.03	16.20	1,230	60.92	17.56
1W0X1	1,411	80.54	10.99	1,418	77.03	15.87	1,418	79.50	12.98	1,418	81.19	9.74	1,418	81.42	12.06
2S0X1	3,296	55.98	14.84	3,435	46.57	20.04	3,435	61.99	13.78	3,435	54.43	17.32	3,435	54.74	18.51
2W1X1	3,224	63.95	17.04	3,298	64.84	21.20	3,298	66.35	16.26	3,298	63.40	18.30	3,298	69.72	15.90
3E7X1	1,479	63.46	15.77	1,506	64.23	20.36	1,506	64.71	15.48	1,506	64.45	16.03	1,506	66.65	17.08
3P0X1	11,662	61.30	15.74	11,999	58.38	20.23	11,999	63.32	15.49	11,999	61.27	16.42	11,999	62.31	18.17
4N0X1	2,886	66.86	14.63	2,943	57.39	20.20	2,943	69.00	15.16	2,943	66.30	14.67	2,943	65.47	17.51

Exhibit A8. Distributions of predictor-ASVAB correlations across all AFSC, cycles, and grades.

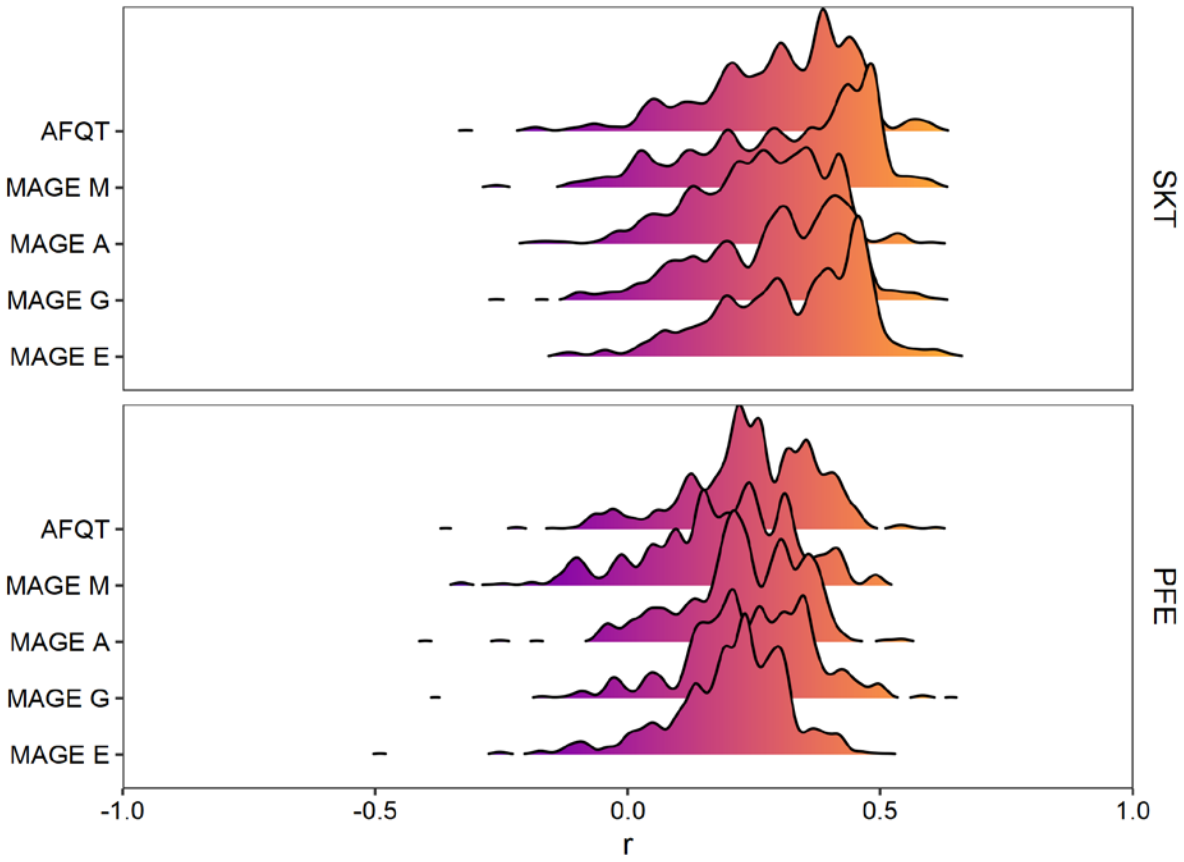


Exhibit A9. Distributions of predictor-ASVAB correlations across all AFSC, cycles, and grades, corrected for multivariate range restriction.

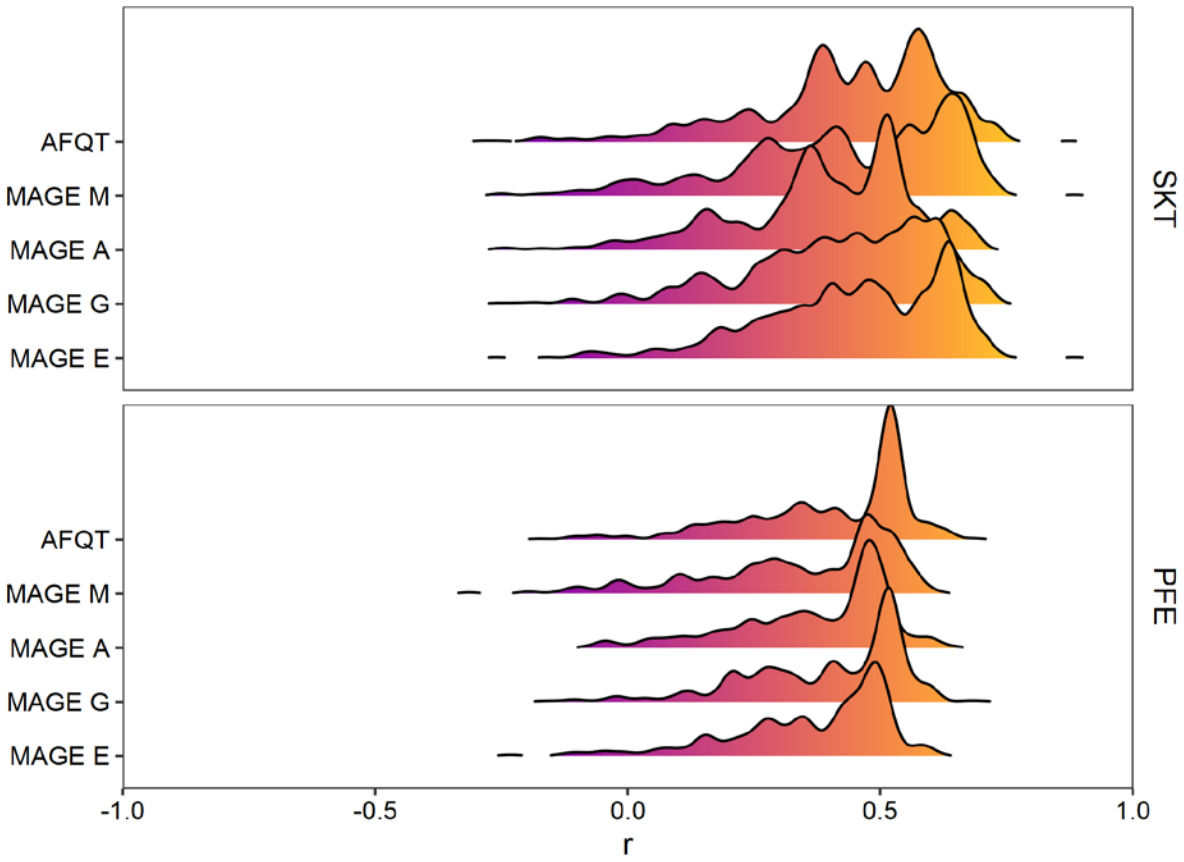


Exhibit A10. Distributions of ASVAB-outcome correlations across all AFSC, cycles, and grades.

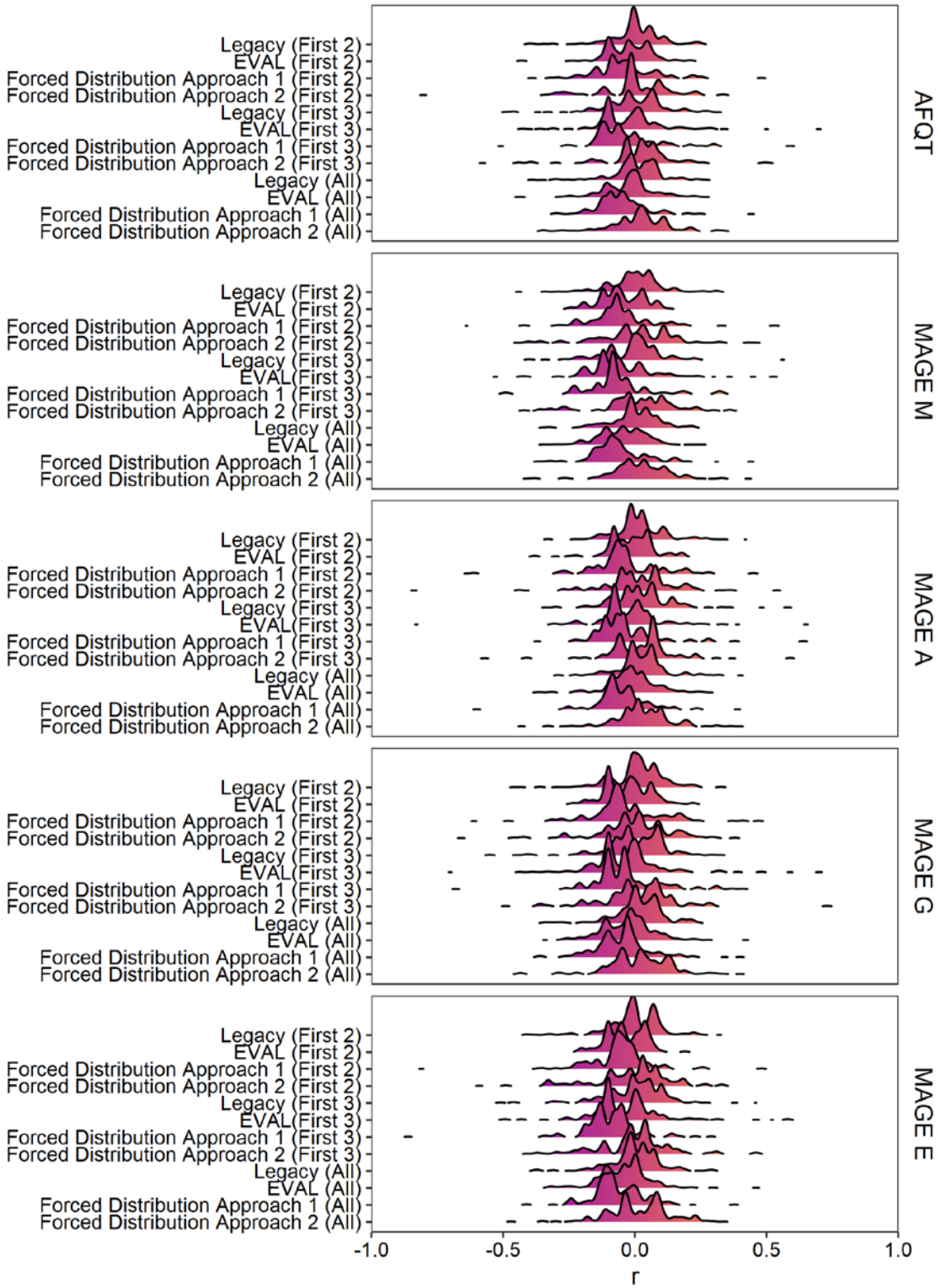
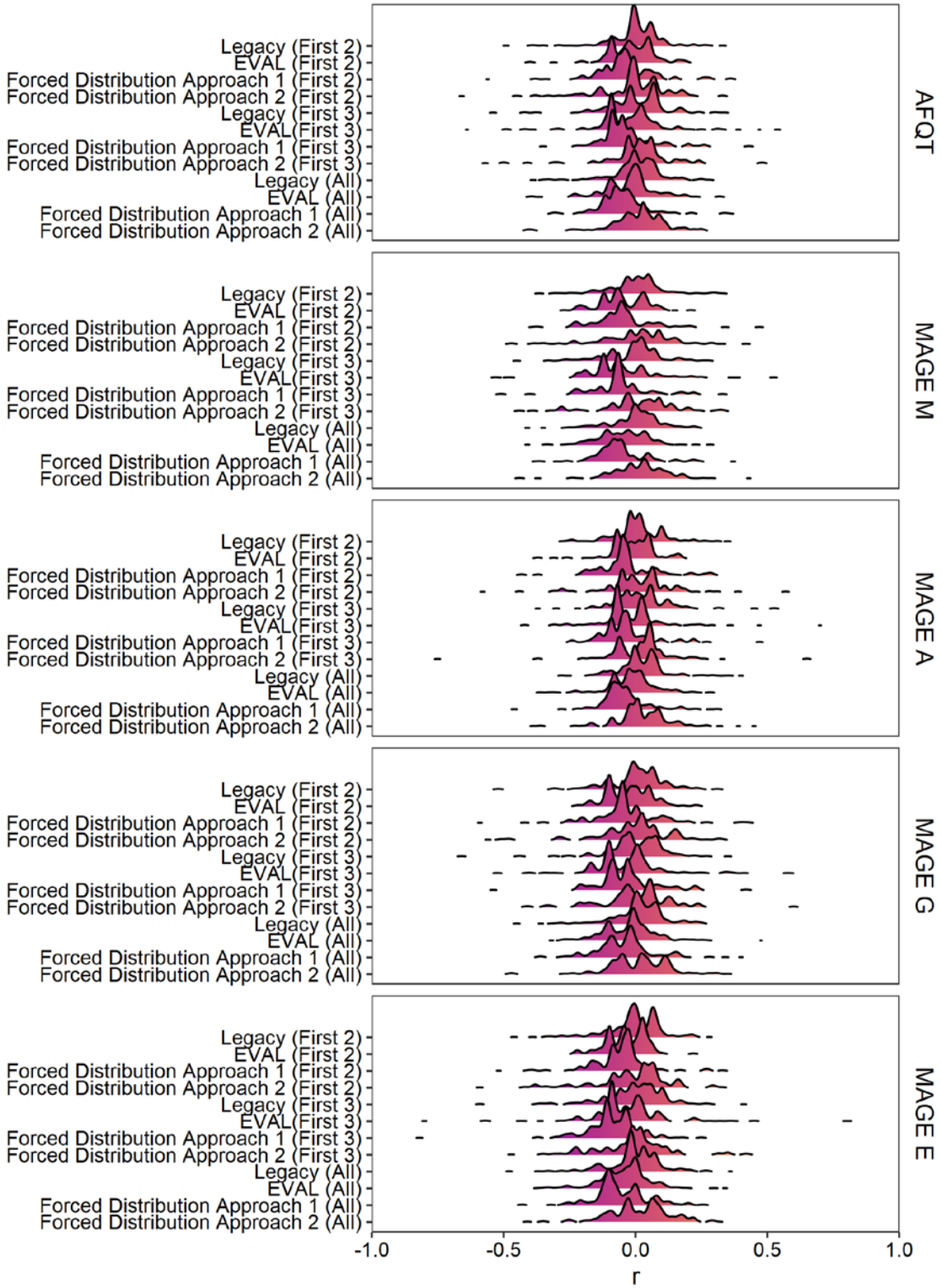


Exhibit A11. Distributions of ASVAB-outcome correlations across all AFSC, cycles, and grades, corrected for multivariate range restriction.



APPENDIX B – ADDITIONAL TASK 2 RESULTS

Table B1. Linear Summary Model of Differences between Item Difficulties (p Values) for First-Time Examinees and Repeat Examinees with Item-Exposure Moderation Effect

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	SE	p	95% CILL	95% CIUL
Model 1b: Overall Difference Moderated by Item-Exposure										
Intercept	81	3,926	564,971	240,323	324,648	-0.02	0.01	.00	-0.04	-0.01
Exposure	66	4,334	415,937	121,498	294,439	0.00	0.00	.01	0.00	0.00
Model 2b: Test-Type Differences Moderated by Item-Exposure										
Main Effects for Unexposed Items										
PFE	20	1,626	478,559	205,138	273,421	-0.01	0.01	.48	-0.04	0.02
SKT	61	2,300	86,412	35,185	51,227	-0.03	0.01	.00	-0.04	-0.01
Item-Exposure Moderator Effects										
PFE	11	372	333,833	92,016	241,817	0.00	0.00	.16	0.00	0.00
SKT	55	3,962	82,104	29,482	52,622	0.00	0.00	.03	0.00	0.01
Model 3b: PFE and SKT Test Differences Moderated by Item-Exposure										
Main Effects for Unexposed Items										
00035	8	670	222,491	126,956	95,535	-0.01	0.02	.71	-0.05	0.03
00036	7	569	173,263	62,902	110,361	-0.01	0.02	.67	-0.05	0.03
00037	5	387	82,805	15,280	67,525	-0.02	0.03	.55	-0.07	0.04
1C151	4	162	2,891	1,443	1,448	0.00	0.03	.90	-0.06	0.05
1C171	2	68	1,153	404	749	-0.01	0.04	.84	-0.09	0.08
1P051	5	295	2,051	945	1,106	-0.01	0.03	.81	-0.06	0.05
1P071	2	135	1,046	339	707	-0.01	0.04	.86	-0.09	0.08
1W051	5	201	2,129	1,156	973	-0.03	0.03	.32	-0.08	0.03
1W071	4	72	2,202	493	1,709	-0.02	0.03	.52	-0.08	0.04
2S051	5	193	5,665	2,614	3,051	0.01	0.03	.58	-0.04	0.07
2S071	3	71	5,154	1,140	4,014	-0.08	0.03	.03	-0.14	-0.01
2W151	5	186	5,491	2,651	2,840	-0.02	0.03	.44	-0.07	0.03
2W171	3	112	4,778	1,050	3,728	-0.07	0.03	.04	-0.14	0.00
3E751	5	196	3,087	1,500	1,587	0.00	0.03	.89	-0.06	0.05
3E771	2	59	1,212	415	797	-0.03	0.04	.52	-0.11	0.06
3P051	5	220	26,951	13,871	13,080	-0.01	0.03	.74	-0.06	0.04
3P071	3	94	13,533	3,281	10,252	-0.17	0.03	.00	-0.23	-0.11
4N051	5	166	5,414	2,958	2,456	0.00	0.03	.91	-0.06	0.05
4N071	3	70	3,655	925	2,730	-0.02	0.03	.60	-0.09	0.05

(Table continues.)

Table B1. (Continued)

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	SE	p	95% CILL	95% CIUL
Model 3b (Cont.): PFE and SKT Test Differences Moderated by Item-Exposure										
Item-Exposure Moderator Effects										
00035	3	130	132,887	62,443	70,444	0.00	0.00	.95	0.00	0.00
00036	3	129	118,141	14,293	103,848	0.01	0.00	.03	0.00	0.01
00037	5	113	82,805	15,280	67,525	0.00	0.00	.30	0.00	0.01
1C151	4	338	3,076	1,384	1,692	0.00	0.01	.84	-0.02	0.01
1C171	2	131	1,153	404	749	0.00	0.01	.88	-0.02	0.02
1P051	4	204	1,684	743	941	0.01	0.01	.09	0.00	0.02
1P071	2	52	1,046	339	707	0.01	0.01	.64	-0.02	0.03
1W051	4	299	1,814	948	866	0.02	0.01	.00	0.01	0.03
1W071	4	316	2,202	493	1,709	0.00	0.01	.92	-0.02	0.02
2S051	4	307	4,498	1,960	2,538	0.00	0.01	.89	-0.01	0.01
2S071	3	227	5,154	1,140	4,014	-0.01	0.01	.27	-0.02	0.01
2W151	4	314	4,444	2,026	2,418	0.01	0.01	.04	0.00	0.02
2W171	3	188	4,778	1,050	3,728	0.00	0.01	.68	-0.01	0.01
3E751	4	304	2,519	1,181	1,338	0.01	0.01	.24	0.00	0.02
3E771	2	136	1,212	415	797	-0.01	0.01	.39	-0.03	0.01
3P051	4	280	22,014	10,731	11,283	0.00	0.00	.43	0.00	0.01
3P071	4	303	18,311	3,322	14,989	0.00	0.00	.56	-0.01	0.01
4N051	4	334	4,544	2,421	2,123	0.00	0.01	.89	-0.01	0.01
4N071	3	229	3,655	925	2,730	0.01	0.01	.10	0.00	0.02

Note. Term = Label for effect tested in the regression model; k_{Admin} = Number of test administrations that contributed to the effect; k_{Item} = Number of items that contributed to the effect; N = Total size of examinee sample; $n_{Firsttime}$ = Size of first-time examinee sample; n_{Repeat} = Size of repeat examinee sample; b = Regression coefficient indicating the average effect at a given level of aggregation; SE = Standard error of b; p = p value of for the significance test for b; 95% CILL = Lower bound of the 95% confidence interval for b; 95% CIUL = Upper bound of the 95% confidence interval for b. A positive b coefficient indicates that the average item statistic for repeat examinees was larger than for first-time examinees. Moderator effects for item exposure are additive terms that reflect the average difference between exposed items and new items.

Table B2. Linear Summary Model of Differences between Corrected Item-Total Correlations for First-Time Examinees and Repeat Examinees with Item-Exposure Moderation Effect

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	SE	p	95% CILL	95% CIUL
Model 1b: Overall Difference Moderated by Item-Exposure										
Intercept	81	3,926	564,971	240,323	324,648	0.00	.01	0.98	-0.01	0.01
Exposure	66	4,334	415,937	121,498	294,439	0.01	.00	0.00	0.00	0.01
Model 2b: Test-Type Differences Moderated by Item-Exposure										
Main Effects for Unexposed Items										
PFE	20	1,626	478,559	205,138	273,421	-0.02	.01	0.11	-0.04	0.00
SKT	61	2,300	86,412	35,185	51,227	0.01	.01	0.29	-0.01	0.02
Item-Exposure Moderator Effects										
PFE	11	372	333,833	92,016	241,817	0.01	.00	0.00	0.00	0.01
SKT	55	3,962	82,104	29,482	52,622	0.00	.00	0.10	0.00	0.01
Model 3b: PFE and SKT Test Differences Moderated by Item-Exposure										
Main Effects for Unexposed Items										
00035	8	670	222,491	126,956	95,535	-0.03	.02	0.07	-0.06	0.00
00036	7	569	173,263	62,902	110,361	-0.01	.02	0.73	-0.04	0.03
00037	5	387	82,805	15,280	67,525	-0.01	.02	0.46	-0.05	0.02
1C151	4	162	2,891	1,443	1,448	0.01	.02	0.58	-0.03	0.05
1C171	2	68	1,153	404	749	0.02	.03	0.64	-0.05	0.08
1P051	5	295	2,051	945	1,106	-0.01	.02	0.66	-0.05	0.03
1P071	2	135	1,046	339	707	0.02	.03	0.52	-0.04	0.08
1W051	5	201	2,129	1,156	973	-0.02	.02	0.46	-0.06	0.03
1W071	4	72	2,202	493	1,709	0.02	.03	0.44	-0.03	0.07
2S051	5	193	5,665	2,614	3,051	-0.01	.02	0.73	-0.05	0.03
2S071	3	71	5,154	1,140	4,014	0.07	.03	0.02	0.01	0.12
2W151	5	186	5,491	2,651	2,840	-0.02	.02	0.31	-0.06	0.02
2W171	3	112	4,778	1,050	3,728	0.07	.03	0.01	0.02	0.12
3E751	5	196	3,087	1,500	1,587	-0.01	.02	0.58	-0.05	0.03
3E771	2	59	1,212	415	797	0.04	.03	0.19	-0.02	0.11
3P051	5	220	26,951	13,871	13,080	-0.03	.02	0.19	-0.06	0.01
3P071	3	94	13,533	3,281	10,252	0.09	.02	0.00	0.05	0.13
4N051	5	166	5,414	2,958	2,456	-0.01	.02	0.51	-0.05	0.03
4N071	3	70	3,655	925	2,730	0.00	.03	0.92	-0.05	0.05

(Table continues.)

Table B2 (Continued)

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	SE	p	95% CILL	95% CIUL
Model 3b (Cont.): PFE and SKT Test Differences Moderated by Item-Exposure										
Item-Exposure Moderator Effects										
00035	3	130	132,887	62,443	70,444	0.01	0.00	.00	0.01	0.02
00036	3	129	118,141	14,293	103,848	-0.01	0.00	.01	-0.02	0.00
00037	5	113	82,805	15,280	67,525	0.01	0.00	.04	0.00	0.02
1C151	4	338	3,076	1,384	1,692	-0.02	0.01	.18	-0.04	0.01
1C171	2	131	1,153	404	749	0.00	0.02	.95	-0.03	0.03
1P051	4	204	1,684	743	941	0.00	0.01	.71	-0.03	0.02
1P071	2	52	1,046	339	707	0.00	0.02	.84	-0.04	0.04
1W051	4	299	1,814	948	866	0.01	0.01	.59	-0.02	0.03
1W071	4	316	2,202	493	1,709	0.00	0.02	.82	-0.03	0.04
2S051	4	307	4,498	1,960	2,538	0.00	0.01	.82	-0.02	0.02
2S071	3	227	5,154	1,140	4,014	0.00	0.01	.87	-0.02	0.02
2W151	4	314	4,444	2,026	2,418	0.01	0.01	.22	-0.01	0.03
2W171	3	188	4,778	1,050	3,728	0.00	0.01	.70	-0.02	0.02
3E751	4	304	2,519	1,181	1,338	0.00	0.01	.65	-0.02	0.03
3E771	2	136	1,212	415	797	0.00	0.02	.94	-0.03	0.03
3P051	4	280	22,014	10,731	11,283	0.01	0.00	.04	0.00	0.02
3P071	4	303	18,311	3,322	14,989	0.01	0.01	.29	-0.01	0.02
4N051	4	334	4,544	2,421	2,123	-0.02	0.01	.08	-0.04	0.00
4N071	3	229	3,655	925	2,730	0.01	0.01	.36	-0.01	0.03

Note. Term = Label for effect tested in the regression model; k_{Admin} = Number of test administrations that contributed to the effect; k_{Item} = Number of items that contributed to the effect; N = Total size of examinee sample; $n_{Firsttime}$ = Size of first-time examinee sample; n_{Repeat} = Size of repeat examinee sample; b = Regression coefficient indicating the average effect at a given level of aggregation; SE = Standard error of b; p = p value of for the significance test for b; 95% CILL = Lower bound of the 95% confidence interval for b; 95% CIUL = Upper bound of the 95% confidence interval for b. A positive b coefficient indicates that the average item statistic for repeat examinees was larger than for first-time examinees. Moderator effects for item exposure are additive terms that reflect the average difference between exposed items and new items.

Table B3. Linear Summary Model of Differences between Item-AFQT Correlations for First-Time Examinees and Repeat Examinees with Item-Exposure Moderation Effect

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	SE	p	95% CILL	95% CIUL
Model 1b: Overall Difference Moderated by Item-Exposure										
Intercept	81	3,926	564,971	240,323	324,648	-0.02	0.00	.00	-0.02	-0.01
Exposure	66	4,334	415,937	121,498	294,439	0.00	0.00	.03	-0.01	0.00
Model 2b: Test-Type Differences Moderated by Item-Exposure										
Main Effects for Unexposed Items										
PFE	20	1,626	478,559	205,138	273,421	-0.02	0.01	.00	-0.03	-0.01
SKT	61	2,300	86,412	35,185	51,227	-0.02	0.00	.00	-0.03	-0.01
Item-Exposure Moderator Effects										
PFE	11	372	333,833	92,016	241,817	0.00	0.00	.02	-0.01	0.00
SKT	55	3,962	82,104	29,482	52,622	0.00	0.00	.70	-0.01	0.00
Model 3b: PFE and SKT Test Differences Moderated by Item-Exposure										
Main Effects for Unexposed Items										
00035	8	670	222,491	126,956	95,535	-0.03	0.01	.00	-0.04	-0.02
00036	7	569	173,263	62,902	110,361	-0.01	0.01	.13	-0.02	0.00
00037	5	387	82,805	15,280	67,525	-0.02	0.01	.02	-0.03	0.00
1C151	4	162	2,891	1,443	1,448	-0.03	0.01	.03	-0.05	0.00
1C171	2	68	1,153	404	749	-0.07	0.02	.00	-0.11	-0.03
1P051	5	295	2,051	945	1,106	-0.03	0.01	.01	-0.05	-0.01
1P071	2	135	1,046	339	707	-0.05	0.02	.00	-0.09	-0.02
1W051	5	201	2,129	1,156	973	-0.02	0.01	.08	-0.05	0.00
1W071	4	72	2,202	493	1,709	0.02	0.02	.33	-0.02	0.07
2S051	5	193	5,665	2,614	3,051	-0.02	0.01	.06	-0.04	0.00
2S071	3	71	5,154	1,140	4,014	0.01	0.02	.53	-0.02	0.04
2W151	5	186	5,491	2,651	2,840	-0.03	0.01	.02	-0.05	0.00
2W171	3	112	4,778	1,050	3,728	0.02	0.01	.25	-0.01	0.04
3E751	5	196	3,087	1,500	1,587	-0.03	0.01	.01	-0.06	-0.01
3E771	2	59	1,212	415	797	-0.04	0.03	.11	-0.09	0.01
3P051	5	220	26,951	13,871	13,080	-0.02	0.01	.01	-0.04	-0.01
3P071	3	94	13,533	3,281	10,252	0.00	0.01	.99	-0.02	0.02
4N051	5	166	5,414	2,958	2,456	-0.03	0.01	.02	-0.05	0.00
4N071	3	70	3,655	925	2,730	-0.03	0.02	.06	-0.07	0.00

(Table continues.)

Table B3 (Continued)

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	SE	p	95% CILL	95% CIUL
Model 3b (Cont.): PFE and SKT Test Differences Moderated by Item-Exposure										
Item-Exposure Moderator Effects										
00035	3	130	132,887	62,443	70,444	-0.01	0.00	.00	-0.01	0.00
00036	3	129	118,141	14,293	103,848	0.00	0.00	.42	-0.01	0.00
00037	5	113	82,805	15,280	67,525	0.00	0.00	.60	-0.01	0.01
1C151	4	338	3,076	1,384	1,692	0.01	0.01	.30	-0.01	0.04
1C171	2	131	1,153	404	749	0.00	0.02	.86	-0.04	0.05
1P051	4	204	1,684	743	941	-0.02	0.01	.20	-0.04	0.01
1P071	2	52	1,046	339	707	-0.04	0.03	.19	-0.09	0.02
1W051	4	299	1,814	948	866	0.00	0.01	.79	-0.02	0.03
1W071	4	316	2,202	493	1,709	0.01	0.02	.63	-0.03	0.06
2S051	4	307	4,498	1,960	2,538	-0.01	0.01	.22	-0.03	0.01
2S071	3	227	5,154	1,140	4,014	0.00	0.01	.93	-0.03	0.03
2W151	4	314	4,444	2,026	2,418	-0.01	0.01	.31	-0.03	0.01
2W171	3	188	4,778	1,050	3,728	0.01	0.01	.49	-0.02	0.03
3E751	4	304	2,519	1,181	1,338	0.00	0.01	.88	-0.02	0.03
3E771	2	136	1,212	415	797	0.08	0.03	.00	0.02	0.13
3P051	4	280	22,014	10,731	11,283	-0.01	0.00	.07	-0.02	0.00
3P071	4	303	18,311	3,322	14,989	0.01	0.01	.09	0.00	0.03
4N051	4	334	4,544	2,421	2,123	-0.02	0.01	.18	-0.04	0.01
4N071	3	229	3,655	925	2,730	0.02	0.02	.21	-0.01	0.05

Note. Term = Label for effect tested in the regression model; k_{Admin} = Number of test administrations that contributed to the effect; k_{Item} = Number of items that contributed to the effect; N = Total size of examinee sample; $n_{Firsttime}$ = Size of first-time examinee sample; n_{Repeat} = Size of repeat examinee sample; b = Regression coefficient indicating the average effect at a given level of aggregation; SE = Standard error of b; p = p value of for the significance test for b; 95% CILL = Lower bound of the 95% confidence interval for b; 95% CIUL = Upper bound of the 95% confidence interval for b. A positive b coefficient indicates that the average item statistic for repeat examinees was larger than for first-time examinees. Moderator effects for item exposure are additive terms that reflect the average difference between exposed items and new items.

Table B4. Linear Summary Model of Differential Item Functioning (DIF) Odds Ratios Comparing First-Time Examinees and Repeat Examinees with Item-Exposure Moderation Effect

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	p	95% CILL	95% CIUL
Model 1b: Overall DIF Moderated by Item-Exposure									
Intercept	65	2,949	331,297	152,909	178,388	1.02	.07	1.00	1.04
Exposure	53	3,307	252,521	108,292	144,229	0.02	.00	0.01	0.03
Model 2b: Test-Type DIF Moderated by Item-Exposure									
Main Effects for Unexposed Items									
PFE	12	943	273,362	124,887	148,475	1.04	.06	1.00	1.08
SKT	53	2,006	57,935	28,022	29,913	1.01	.45	1.00	1.03
Item-Exposure Moderator Effects									
PFE	7	179	203,205	85,596	117,609	0.01	.10	0.00	0.03
SKT	46	3,128	49,316	22,696	26,620	0.02	.01	0.00	0.05
Model 3b: PFE and SKT Test DIF Moderated by Item-Exposure									
Main Effects for Unexposed Items									
00035	6	435	184,036	89,620	94,416	1.04	.14	0.99	1.09
00036	3	251	55,955	22,074	33,881	1.09	.02	1.02	1.17
00037	3	257	33,371	13,193	20,178	0.99	.89	0.92	1.07
1C151	4	155	2,854	1,410	1,444	0.97	.46	0.89	1.05
1C171	1	47	178	127	51	1.01	.93	0.83	1.23
1P051	5	271	1,986	893	1,093	0.98	.60	0.91	1.05
1P071	1	80	187	122	65	1.10	.27	0.93	1.31
1W051	5	189	2,082	1,117	965	0.93	.07	0.86	1.01
1W071	3	55	629	178	451	1.03	.70	0.88	1.20
2S051	5	179	5,477	2,464	3,013	1.10	.01	1.03	1.18
2S071	1	54	585	358	227	1.02	.80	0.88	1.18
2W151	5	179	5,320	2,511	2,809	0.93	.04	0.87	0.99
2W171	2	81	1,033	368	665	1.12	.06	1.00	1.25
3E751	5	184	3,005	1,426	1,579	1.03	.47	0.95	1.11
3E771	2	49	361	136	225	1.07	.45	0.90	1.28
3P051	5	204	25,969	13,038	12,931	1.02	.55	0.96	1.08
3P071	2	72	2,248	782	1,466	0.96	.50	0.87	1.07
4N051	5	155	5,280	2,832	2,448	1.03	.46	0.96	1.11
4N071	2	52	741	260	481	0.92	.24	0.80	1.06

(Table continues.)

Table B4 (Continued)

Term	k_{Admin}	k_{Item}	N	$n_{Firsttime}$	n_{Repeat}	b	p	95% CILL	95% CIUL
Model 3b (Cont.): PFE and SKT Test DIF Moderated by Item-Exposure									
Item-Exposure Moderator Effects									
00035	3	123	128,801	59,025	69,776	0.02	.05	0.00	0.04
00036	1	27	41,033	13,378	27,655	0.00	.90	-0.05	0.05
00037	3	29	33,371	13,193	20,178	0.00	.97	-0.04	0.05
1C151	4	335	3,032	1,349	1,683	0.01	.85	-0.06	0.10
1C171	1	47	178	127	51	0.02	.83	-0.15	0.35
1P051	4	197	1,636	707	929	0.07	.09	-0.01	0.16
1P071	1	17	187	122	65	0.16	.36	-0.13	0.67
1W051	4	293	1,777	918	859	0.08	.04	0.00	0.18
1W071	3	234	629	178	451	0.02	.85	-0.11	0.22
2S051	4	298	4,358	1,853	2,505	0.02	.55	-0.04	0.10
2S071	1	38	585	358	227	0.05	.45	-0.07	0.23
2W151	4	310	4,325	1,933	2,392	0.08	.01	0.02	0.15
2W171	2	102	1,033	368	665	-0.02	.72	-0.11	0.10
3E751	4	300	2,463	1,129	1,334	0.09	.02	0.01	0.18
3E771	2	125	361	136	225	-0.15	.11	-0.26	0.05
3P051	4	270	21,321	10,150	11,171	0.00	.98	-0.03	0.03
3P071	2	104	2,248	782	1,466	-0.01	.84	-0.07	0.07
4N051	4	323	4,442	2,326	2,116	0.02	.48	-0.04	0.10
4N071	2	135	741	260	481	0.13	.04	0.00	0.32

Note. Term = Label for effect tested in the regression model; k_{Admin} = Number of test administrations that contributed to the effect; k_{Item} = Number of items that contributed to the effect; N = Total size of examinee sample; $n_{Firsttime}$ = Size of first-time examinee sample; n_{Repeat} = Size of repeat examinee sample; b = Regression coefficient indicating the average effect at a given level of aggregation; SE = Standard error of b; $p = p$ value of for the significance test for b; 95% CILL = Lower bound of the 95% confidence interval for b; 95% CIUL = Upper bound of the 95% confidence interval for b. A b coefficient larger than 1 indicates that, on average, repeat examinees were more likely to get items correct than were first-time examinees with similar AFQT scores. Moderator effects for item exposure are additive terms that reflect the average difference between exposed items and new items.

Table B5. Meta-Analyses of Path-Model Coefficients for Female-Male Differences in PFE/SKT Test Scores with Time in Grade (TIG) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	99	525,201	106,233	418,968	−.55	(−0.79, −0.31)	−.55	(−0.79, −0.32)	.00	(−0.01, 0.01)	
PFE	23	458,328	90,735	367,593	−.32	(−0.65, 0.02)	−.32	(−0.66, 0.02)	.00	(−0.00, 0.01)	
00035	9	207,965	40,606	167,359	−.15	(−0.65, 0.35)	−.15	(−0.66, 0.35)	.00	(−0.00, 0.01)	
00036	7	154,934	31,081	123,853	−.63	(−1.06, −0.20)	−.64	(−1.07, −0.20)	.01	(−0.00, 0.02)	
00037	7	95,429	19,048	76,381	−.17	(−1.35, 1.02)	−.16	(−1.35, 1.02)	−.00	(−0.03, 0.03)	
SKT	76	66,873	15,498	51,375	−2.14	(−2.60, −1.69)	−2.12	(−2.57, −1.66)	−.03	(−0.05, −0.00)	X
1C151	6	3,854	588	3,266	−1.28	(−1.84, −0.72)	−1.31	(−1.86, −0.76)	.04	(−0.06, 0.13)	
1C171	5	1,057	209	848	−.14	(−2.82, 2.54)	−.28	(−2.95, 2.38)	.14	(−0.15, 0.43)	
1P051	6	2,044	344	1,700	−.93	(−1.95, 0.09)	−.90	(−1.90, 0.09)	−.03	(−0.10, 0.04)	
1P071	5	949	202	747	−2.33	(−3.96, −0.70)	−2.30	(−4.05, −0.54)	−.04	(−0.16, 0.09)	
1W051	6	2,079	518	1,561	−3.83	(−5.27, −2.39)	−3.83	(−5.24, −2.42)	.00	(−0.15, 0.16)	
1W071	5	970	149	821	−4.02	(−6.05, −2.00)	−4.02	(−6.08, −1.95)	−.01	(−0.58, 0.57)	
2S051	6	5,750	2,024	3,726	.37	(−0.34, 1.09)	.38	(−0.32, 1.09)	−.01	(−0.05, 0.03)	
2S071	5	2,802	1,265	1,537	1.12	(0.20, 2.05)	1.11	(0.18, 2.05)	.01	(−0.02, 0.03)	
2W151	6	5,598	761	4,837	−5.25	(−6.12, −4.37)	−5.19	(−6.12, −4.26)	−.05	(−0.27, 0.16)	
2W171	4	2,121	146	1,975	−4.89	(−8.10, −1.68)	−4.87	(−8.09, −1.65)	−.02	(−0.04, −0.00)	X
3P051	6	26,460	4,784	21,676	−3.19	(−3.79, −2.59)	−3.12	(−3.74, −2.50)	−.07	(−0.18, 0.04)	
3P071	5	5,820	555	5,265	−4.43	(−5.75, −3.12)	−4.39	(−5.66, −3.12)	−.04	(−0.13, 0.05)	
4N051	6	5,424	2,944	2,480	−1.65	(−2.44, −0.86)	−1.66	(−2.49, −0.83)	.01	(−0.04, 0.06)	
4N071	5	1,945	1,009	936	.43	(0.09, 0.77)	.44	(0.11, 0.76)	−.01	(−0.04, 0.02)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B6. Meta-Analyses of Path-Model Coefficients for Female-Male Differences in PFE/SKT Test Scores with Time in Service (TIS) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	99	525,201	106,233	418,968	−.55	(−0.79, −0.31)	−.55	(−0.79, −0.32)	.00	(−0.03, 0.03)	
PFE	23	458,328	90,735	367,593	−.32	(−0.65, 0.02)	−.31	(−0.64, 0.03)	−.01	(−0.06, 0.04)	
00035	9	207,965	40,606	167,359	−.15	(−0.65, 0.35)	−.16	(−0.67, 0.35)	.01	(−0.01, 0.03)	
00036	7	154,934	31,081	123,853	−.63	(−1.06, −0.20)	−.61	(−1.03, −0.18)	−.02	(−0.05, 0.00)	
00037	7	95,429	19,048	76,381	−.17	(−1.35, 1.02)	−.13	(−1.28, 1.03)	−.04	(−0.30, 0.22)	
SKT	76	66,873	15,498	51,375	−2.14	(−2.60, −1.69)	−2.24	(−2.69, −1.79)	.09	(0.02, 0.16)	X
1C151	6	3,854	588	3,266	−1.28	(−1.84, −0.72)	−1.36	(−1.92, −0.81)	.08	(−0.01, 0.18)	
1C171	5	1,057	209	848	−.14	(−2.82, 2.54)	−.59	(−2.91, 1.73)	.45	(−0.28, 1.18)	
1P051	6	2,044	344	1,700	−.93	(−1.95, 0.09)	−.91	(−1.90, 0.07)	−.02	(−0.10, 0.06)	
1P071	5	949	202	747	−2.33	(−3.96, −0.70)	−2.64	(−4.59, −0.69)	.31	(−0.24, 0.86)	
1W051	6	2,079	518	1,561	−3.83	(−5.27, −2.39)	−3.86	(−5.29, −2.42)	.03	(−0.12, 0.17)	
1W071	5	970	149	821	−4.02	(−6.05, −2.00)	−4.61	(−6.65, −2.57)	.59	(−0.24, 1.42)	
2S051	6	5,750	2,024	3,726	.37	(−0.34, 1.09)	.40	(−0.31, 1.11)	−.03	(−0.06, 0.01)	
2S071	5	2,802	1,265	1,537	1.12	(0.20, 2.05)	.44	(−0.43, 1.32)	.68	(0.18, 1.17)	X
2W151	6	5,598	761	4,837	−5.25	(−6.12, −4.37)	−5.26	(−6.20, −4.33)	.02	(−0.19, 0.23)	
2W171	4	2,121	146	1,975	−4.89	(−8.10, −1.68)	−6.16	(−9.11, −3.21)	1.27	(0.40, 2.14)	X
3P051	6	26,460	4,784	21,676	−3.19	(−3.79, −2.59)	−3.14	(−3.75, −2.54)	−.05	(−0.13, 0.04)	
3P071	5	5,820	555	5,265	−4.43	(−5.75, −3.12)	−4.73	(−6.08, −3.39)	.30	(0.16, 0.44)	X
4N051	6	5,424	2,944	2,480	−1.65	(−2.44, −0.86)	−1.67	(−2.53, −0.81)	.02	(−0.06, 0.10)	
4N071	5	1,945	1,009	936	.43	(0.09, 0.77)	.16	(−0.25, 0.58)	.26	(−0.01, 0.54)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B7. Meta-Analyses of Path-Model Coefficients for Black-White Differences in PFE/SKT Test Scores with Time in Grade (TIG) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	106	372,092	75,024	297,068	−2.05	(−2.26, −1.85)	−2.10	(−2.31, −1.90)	.05	(0.03, 0.07)	X
PFE	23	325,414	64,010	261,404	−1.86	(−2.21, −1.51)	−1.92	(−2.28, −1.56)	.06	(0.02, 0.10)	X
00035	9	139,922	26,776	113,146	−1.93	(−2.43, −1.42)	−1.96	(−2.45, −1.47)	.03	(−0.02, 0.09)	
00036	7	116,630	22,706	93,924	−2.16	(−2.51, −1.82)	−2.30	(−2.64, −1.97)	.14	(0.08, 0.20)	X
00037	7	68,862	14,528	54,334	−1.25	(−2.44, −0.05)	−1.22	(−2.39, −0.05)	−.02	(−0.05, 0.00)	
SKT	83	46,678	11,014	35,664	−3.30	(−3.64, −2.96)	−3.29	(−3.63, −2.95)	−.01	(−0.04, 0.02)	
1C151	6	2,581	330	2,251	−2.11	(−2.90, −1.31)	−2.03	(−2.72, −1.34)	−.07	(−0.29, 0.14)	
1C171	5	844	154	690	−2.95	(−5.02, −0.89)	−2.87	(−4.69, −1.05)	−.08	(−0.45, 0.29)	
1P051	6	1,391	274	1,117	−2.37	(−3.09, −1.64)	−2.39	(−3.09, −1.68)	.02	(−0.04, 0.08)	
1P071	5	694	207	487	−3.31	(−5.31, −1.31)	−3.28	(−5.37, −1.19)	−.03	(−0.18, 0.12)	
1W051	4	1,302	149	1,153	−3.78	(−5.39, −2.18)	−3.66	(−5.42, −1.89)	−.13	(−0.43, 0.18)	
1W071	3	523	67	456	−4.98	(−5.04, −4.91)	−4.98	(−5.86, −4.10)	.00	(−0.87, 0.87)	
2S051	6	3,617	1,478	2,139	−1.03	(−1.96, −0.09)	−1.14	(−2.24, −0.05)	.11	(−0.10, 0.32)	
2S071	5	1,757	979	778	−.30	(−0.49, −0.11)	−.30	(−0.47, −0.12)	−.00	(−0.05, 0.05)	
2W151	6	3,743	704	3,039	−4.26	(−5.35, −3.16)	−4.07	(−5.06, −3.09)	−.18	(−0.40, 0.04)	
2W171	5	1,894	431	1,463	−3.80	(−5.28, −2.32)	−3.82	(−5.30, −2.33)	.02	(−0.03, 0.06)	
3E751	6	2,025	319	1,706	−4.82	(−5.69, −3.94)	−4.79	(−5.74, −3.84)	−.03	(−0.24, 0.19)	
3E771	4	616	106	510	−5.14	(−6.90, −3.39)	−5.23	(−6.75, −3.70)	.08	(−0.28, 0.45)	
3P051	6	16,635	3,724	12,911	−4.20	(−4.64, −3.75)	−4.18	(−4.59, −3.77)	−.02	(−0.08, 0.05)	
3P071	5	4,202	847	3,355	−4.17	(−5.84, −2.50)	−4.17	(−5.78, −2.56)	.00	(−0.07, 0.08)	
4N051	6	3,617	840	2,777	−2.94	(−3.99, −1.89)	−2.96	(−3.97, −1.96)	.03	(−0.08, 0.13)	
4N071	5	1,237	405	832	−1.34	(−2.71, 0.03)	−1.37	(−2.74, −0.00)	.03	(−0.02, 0.09)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B8. Meta-Analyses of Path-Model Coefficients for Black-White Differences in PFE/SKT Test Scores with Time in Service (TIS) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	106	372,092	75,024	297,068	−2.05	(−2.26, −1.85)	−1.82	(−2.05, −1.60)	−.23	(−0.31, −0.15)	X
PFE	23	325,414	64,010	261,404	−1.86	(−2.21, −1.51)	−1.62	(−2.04, −1.20)	−.24	(−0.43, −0.05)	X
00035	9	139,922	26,776	113,146	−1.93	(−2.43, −1.42)	−1.94	(−2.44, −1.45)	.01	(−0.04, 0.07)	
00036	7	116,630	22,706	93,924	−2.16	(−2.51, −1.82)	−2.08	(−2.39, −1.77)	−.08	(−0.21, 0.04)	
00037	7	68,862	14,528	54,334	−1.25	(−2.44, −0.05)	−.28	(−1.22, 0.66)	−.97	(−1.28, −0.65)	X
SKT	83	46,678	11,014	35,664	−3.30	(−3.64, −2.96)	−3.12	(−3.45, −2.79)	−.18	(−0.26, −0.10)	X
1C151	6	2,581	330	2,251	−2.11	(−2.90, −1.31)	−2.02	(−2.73, −1.31)	−.09	(−0.26, 0.08)	
1C171	5	844	154	690	−2.95	(−5.02, −0.89)	−2.23	(−3.78, −0.68)	−.72	(−1.58, 0.13)	
1P051	6	1,391	274	1,117	−2.37	(−3.09, −1.64)	−2.39	(−3.04, −1.74)	.02	(−0.10, 0.14)	
1P071	5	694	207	487	−3.31	(−5.31, −1.31)	−2.66	(−4.32, −1.01)	−.65	(−1.04, −0.25)	X
1W051	4	1,302	149	1,153	−3.78	(−5.39, −2.18)	−3.58	(−5.10, −2.06)	−.21	(−0.42, 0.01)	
1W071	3	523	67	456	−4.98	(−5.04, −4.91)	−3.27	(−5.31, −1.24)	−1.70	(−3.79, 0.39)	
2S051	6	3,617	1,478	2,139	−1.03	(−1.96, −0.09)	−1.10	(−2.16, −0.04)	.07	(−0.08, 0.22)	
2S071	5	1,757	979	778	−.30	(−0.49, −0.11)	−.08	(−0.76, 0.59)	−.22	(−0.77, 0.34)	
2W151	6	3,743	704	3,039	−4.26	(−5.35, −3.16)	−4.08	(−5.04, −3.12)	−.18	(−0.39, 0.04)	
2W171	5	1,894	431	1,463	−3.80	(−5.28, −2.32)	−3.18	(−4.38, −1.98)	−.62	(−1.02, −0.22)	X
3E751	6	2,025	319	1,706	−4.82	(−5.69, −3.94)	−4.76	(−5.68, −3.84)	−.05	(−0.24, 0.13)	
3E771	4	616	106	510	−5.14	(−6.90, −3.39)	−3.38	(−5.76, −1.01)	−1.76	(−3.56, 0.04)	
3P051	6	16,635	3,724	12,911	−4.20	(−4.64, −3.75)	−4.17	(−4.56, −3.79)	−.02	(−0.11, 0.07)	
3P071	5	4,202	847	3,355	−4.17	(−5.84, −2.50)	−3.53	(−5.25, −1.81)	−.64	(−0.87, −0.42)	X
4N051	6	3,617	840	2,777	−2.94	(−3.99, −1.89)	−2.89	(−3.87, −1.92)	−.04	(−0.16, 0.08)	
4N071	5	1,237	405	832	−1.34	(−2.71, 0.03)	−1.08	(−2.52, 0.35)	−.26	(−0.43, −0.10)	X

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B9. Meta-Analyses of Path-Model Coefficients for Asian-White Differences in PFE/SKT Test Scores with Time in Grade (TIG) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	52	295,487	12,670	282,817	−1.17	(−1.47, −0.88)	−1.19	(−1.49, −0.90)	.02	(0.00, 0.03)	X
PFE	23	273,010	11,606	261,404	−1.19	(−1.58, −0.81)	−1.22	(−1.61, −0.83)	.03	(0.01, 0.04)	X
00035	9	118,252	5,106	113,146	−.73	(−1.49, 0.04)	−.74	(−1.51, 0.03)	.01	(−0.01, 0.03)	
00036	7	98,151	4,227	93,924	−1.45	(−1.84, −1.06)	−1.50	(−1.88, −1.13)	.05	(0.01, 0.10)	X
00037	7	56,607	2,273	54,334	−1.77	(−2.63, −0.91)	−1.76	(−2.62, −0.91)	−.00	(−0.01, 0.00)	
SKT	29	22,477	1,064	21,413	−.95	(−1.80, −0.11)	−.90	(−1.76, −0.05)	−.05	(−0.09, −0.01)	X
2S051	6	2,361	222	2,139	1.35	(−0.60, 3.29)	1.45	(−0.51, 3.41)	−.10	(−0.26, 0.06)	
2S071	2	388	49	339	3.68	(−5.21, 12.57)	3.61	(−4.07, 11.29)	.06	(−1.14, 1.27)	
2W151	5	2,858	130	2,728	−.92	(−2.44, 0.59)	−.78	(−2.22, 0.67)	−.15	(−0.33, 0.03)	
2W171	3	1,101	77	1,024	−4.51	(−7.45, −1.57)	−4.49	(−7.60, −1.39)	−.02	(−0.18, 0.15)	
3P051	5	12,316	319	11,997	−2.57	(−3.31, −1.83)	−2.55	(−3.34, −1.75)	−.02	(−0.11, 0.06)	
4N051	6	2,997	220	2,777	−.66	(−1.36, 0.03)	−.64	(−1.33, 0.06)	−.03	(−0.05, −0.00)	X
4N071	2	456	47	409	−.14	(−15.99, 15.70)	−.13	(−15.87, 15.61)	−.02	(−0.12, 0.09)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B10. Meta-Analyses of Path-Model Coefficients for Asian-White Differences in PFE/SKT Test Scores with Time in Service (TIS) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	52	295,487	12,670	282,817	−1.17	(−1.47, −0.88)	−1.16	(−1.44, −0.87)	−.02	(−0.05, 0.02)	
PFE	23	273,010	11,606	261,404	−1.19	(−1.58, −0.81)	−1.17	(−1.55, −0.79)	−.02	(−0.06, 0.01)	
00035	9	118,252	5,106	113,146	−.73	(−1.49, 0.04)	−.74	(−1.51, 0.03)	.01	(−0.01, 0.03)	
00036	7	98,151	4,227	93,924	−1.45	(−1.84, −1.06)	−1.44	(−1.82, −1.06)	−.01	(−0.04, 0.02)	
00037	7	56,607	2,273	54,334	−1.77	(−2.63, −0.91)	−1.64	(−2.46, −0.82)	−.13	(−0.24, −0.02)	X
SKT	29	22,477	1,064	21,413	−.95	(−1.80, −0.11)	−.98	(−1.80, −0.16)	.03	(−0.11, 0.17)	
2S051	6	2,361	222	2,139	1.35	(−0.60, 3.29)	1.46	(−0.53, 3.45)	−.12	(−0.28, 0.05)	
2S071	2	388	49	339	3.68	(−5.21, 12.57)	2.43	(1.70, 3.16)	1.25	(−6.92, 9.41)	
2W151	5	2,858	130	2,728	−.92	(−2.44, 0.59)	−.80	(−2.24, 0.63)	−.12	(−0.37, 0.13)	
2W171	3	1,101	77	1,024	−4.51	(−7.45, −1.57)	−4.60	(−6.53, −2.67)	.09	(−1.46, 1.64)	
3P051	5	12,316	319	11,997	−2.57	(−3.31, −1.83)	−2.56	(−3.34, −1.78)	−.01	(−0.07, 0.05)	
4N051	6	2,997	220	2,777	−.66	(−1.36, 0.03)	−.63	(−1.32, 0.06)	−.03	(−0.08, 0.01)	
4N071	2	456	47	409	−.14	(−15.99, 15.70)	−.55	(−16.98, 15.87)	.41	(−0.16, 0.99)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B11. Meta-Analyses of Path-Model Coefficients for Pacific Islander-White Differences in PFE/SKT Test Scores with Time in Grade (TIG) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	29	280,183	5,868	274,315	−2.10	(−2.39, −1.81)	−2.12	(−2.40, −1.83)	.02	(−0.01, 0.04)	
PFE	23	267,004	5,600	261,404	−2.10	(−2.41, −1.80)	−2.12	(−2.42, −1.82)	.02	(−0.01, 0.05)	
00035	9	115,370	2,224	113,146	−1.78	(−2.41, −1.14)	−1.82	(−2.45, −1.20)	.05	(−0.01, 0.10)	
00036	7	96,245	2,321	93,924	−2.47	(−2.73, −2.21)	−2.47	(−2.76, −2.18)	−.00	(−0.07, 0.07)	
00037	7	55,389	1,055	54,334	−1.99	(−2.70, −1.29)	−1.99	(−2.70, −1.28)	−.01	(−0.02, 0.01)	
SKT	6	13,179	268	12,911	−2.01	(−3.81, −0.22)	−1.98	(−3.77, −0.20)	−.03	(−0.13, 0.07)	
3P051	6	13,179	268	12,911	−2.01	(−3.81, −0.22)	−1.98	(−3.77, −0.20)	−.03	(−0.13, 0.07)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B12. Meta-Analyses of Path-Model Coefficients for Pacific Islander-White Differences in PFE/SKT Test Scores with Time in Service (TIS) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	29	280,183	5,868	274,315	−2.10	(−2.39, −1.81)	−2.06	(−2.34, −1.78)	−.04	(−0.10, 0.02)	
PFE	23	267,004	5,600	261,404	−2.10	(−2.41, −1.80)	−2.06	(−2.35, −1.77)	−.04	(−0.12, 0.03)	
00035	9	115,370	2,224	113,146	−1.78	(−2.41, −1.14)	−1.80	(−2.43, −1.17)	.03	(−0.02, 0.07)	
00036	7	96,245	2,321	93,924	−2.47	(−2.73, −2.21)	−2.43	(−2.67, −2.20)	−.04	(−0.08, 0.01)	
00037	7	55,389	1,055	54,334	−1.99	(−2.70, −1.29)	−1.78	(−2.41, −1.15)	−.22	(−0.50, 0.07)	
SKT	6	13,179	268	12,911	−2.01	(−3.81, −0.22)	−2.00	(−3.77, −0.23)	−.01	(−0.11, 0.09)	
3P051	6	13,179	268	12,911	−2.01	(−3.81, −0.22)	−2.00	(−3.77, −0.23)	−.01	(−0.11, 0.09)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B13. Meta-Analyses of Path-Model Coefficients for American Indian-White Differences in PFE/SKT Test Scores with Time in Grade (TIG) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	25	270,780	2,780	268,000	−2.02	(−2.51, −1.53)	−2.08	(−2.56, −1.59)	.06	(0.03, 0.08)	X
PFE	21	259,952	2,641	257,311	−1.92	(−2.38, −1.46)	−1.98	(−2.44, −1.52)	.06	(0.04, 0.09)	X
00035	9	114,283	1,137	113,146	−1.94	(−2.53, −1.34)	−2.00	(−2.61, −1.39)	.06	(0.01, 0.11)	X
00036	7	94,936	1,012	93,924	−1.87	(−2.75, −0.99)	−1.96	(−2.84, −1.08)	.09	(0.04, 0.14)	X
00037	5	50,733	492	50,241	−1.97	(−4.09, 0.16)	−1.99	(−4.10, 0.13)	.02	(−0.01, 0.05)	
SKT	4	10,828	139	10,689	−3.92	(−8.05, 0.22)	−3.87	(−7.98, 0.23)	−.04	(−0.17, 0.08)	
3P051	4	10,828	139	10,689	−3.92	(−8.05, 0.22)	−3.87	(−7.98, 0.23)	−.04	(−0.17, 0.08)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B14. Meta-Analyses of Path-Model Coefficients for American Indian-White Differences in PFE/SKT Test Scores with Time in Service (TIS) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	25	270,780	2,780	268,000	−2.02	(−2.51, −1.53)	−1.95	(−2.46, −1.44)	−.06	(−0.14, 0.01)	
PFE	21	259,952	2,641	257,311	−1.92	(−2.38, −1.46)	−1.85	(−2.33, −1.37)	−.07	(−0.15, 0.01)	
00035	9	114,283	1,137	113,146	−1.94	(−2.53, −1.34)	−1.99	(−2.59, −1.38)	.05	(0.00, 0.09)	X
00036	7	94,936	1,012	93,924	−1.87	(−2.75, −0.99)	−1.83	(−2.76, −0.91)	−.04	(−0.09, 0.01)	
00037	5	50,733	492	50,241	−1.97	(−4.09, 0.16)	−1.58	(−3.79, 0.64)	−.39	(−0.54, −0.24)	X
SKT	4	10,828	139	10,689	−3.92	(−8.05, 0.22)	−3.88	(−7.97, 0.21)	−.04	(−0.16, 0.08)	
3P051	4	10,828	139	10,689	−3.92	(−8.05, 0.22)	−3.88	(−7.97, 0.21)	−.04	(−0.16, 0.08)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B15. Meta-Analyses of Path-Model Coefficients for Hispanic-non-Hispanic Differences in PFE/SKT Test Scores with Time in Grade (TIG) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	72	515,904	26,080	489,824	−.89	(−1.05, −0.74)	−.87	(−1.02, −0.72)	−.02	(−0.06, 0.02)	
PFE	23	458,328	23,388	434,940	−.92	(−1.14, −0.70)	−.90	(−1.11, −0.69)	−.02	(−0.09, 0.05)	
00035	9	207,965	5,236	202,729	−.59	(−0.99, −0.20)	−.58	(−0.96, −0.19)	−.02	(−0.06, 0.02)	
00036	7	154,934	9,035	145,899	−1.16	(−1.41, −0.91)	−1.24	(−1.48, −0.99)	.08	(−0.05, 0.20)	
00037	7	95,429	9,117	86,312	−.88	(−1.44, −0.31)	−.76	(−1.19, −0.33)	−.12	(−0.30, 0.06)	
SKT	49	57,576	2,692	54,884	−.66	(−1.06, −0.25)	−.60	(−0.99, −0.21)	−.06	(−0.13, 0.02)	
1C171	3	671	69	602	1.22	(−1.32, 3.75)	1.52	(0.37, 2.67)	−.30	(−2.85, 2.25)	
1P071	1	208	21	187	2.53	(—, —)	2.57	(—, —)	−.03	(—, —)	
2S051	6	5,750	244	5,506	.19	(−2.41, 2.79)	.24	(−2.36, 2.84)	−.05	(−0.13, 0.03)	
2S071	5	2,802	344	2,458	.88	(−0.97, 2.74)	.69	(−0.99, 2.38)	.19	(−0.09, 0.47)	
2W151	6	5,598	165	5,433	−1.32	(−1.91, −0.73)	−.75	(−1.63, 0.13)	−.57	(−0.93, −0.21)	X
2W171	5	2,681	235	2,446	−.32	(−1.34, 0.70)	−.34	(−1.36, 0.69)	.02	(−0.08, 0.13)	
3E771	1	217	20	197	−1.89	(—, —)	−1.33	(—, —)	−.56	(—, —)	
3P051	6	26,460	695	25,765	−1.27	(−2.03, −0.51)	−1.20	(−1.93, −0.47)	−.07	(−0.17, 0.02)	
3P071	5	5,820	522	5,298	−1.33	(−2.46, −0.21)	−1.32	(−2.34, −0.31)	−.01	(−0.14, 0.13)	
4N051	6	5,424	137	5,287	−.06	(−1.61, 1.49)	−.01	(−1.37, 1.35)	−.05	(−0.29, 0.18)	
4N071	5	1,945	240	1,705	−1.25	(−2.22, −0.27)	−1.20	(−2.24, −0.17)	−.04	(−0.17, 0.09)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.

Table B16. Meta-Analyses of Path-Model Coefficients for Hispanic-non-Hispanic Differences in PFE/SKT Test Scores with Time in Service (TIS) as a Mediator

					Total Effect		Direct Effect		Indirect Effect		
Test	***k***	***N***	n_{Focal}	$n_{Referent}$	***Mean***	**95% CI**	***Mean***	**95% CI**	***Mean***	**95% CI**	**Sig. Indirect**
Overall	72	515,904	26,080	489,824	−.89	(−1.05, −0.74)	−.87	(−1.02, −0.72)	−.02	(−0.06, 0.02)	
PFE	23	458,328	23,388	434,940	−.92	(−1.14, −0.70)	−.90	(−1.11, −0.69)	−.02	(−0.09, 0.05)	
00035	9	207,965	5,236	202,729	−.59	(−0.99, −0.20)	−.58	(−0.96, −0.19)	−.02	(−0.06, 0.02)	
00036	7	154,934	9,035	145,899	−1.16	(−1.41, −0.91)	−1.24	(−1.48, −0.99)	.08	(−0.05, 0.20)	
00037	7	95,429	9,117	86,312	−.88	(−1.44, −0.31)	−.76	(−1.19, −0.33)	−.12	(−0.30, 0.06)	
SKT	49	57,576	2,692	54,884	−.66	(−1.06, −0.25)	−.60	(−0.99, −0.21)	−.06	(−0.13, 0.02)	
1C171	3	671	69	602	1.22	(−1.32, 3.75)	1.52	(0.37, 2.67)	−.30	(−2.85, 2.25)	
1P071	1	208	21	187	2.53	(—, —)	2.57	(—, —)	−.03	(—, —)	
2S051	6	5,750	244	5,506	.19	(−2.41, 2.79)	.24	(−2.36, 2.84)	−.05	(−0.13, 0.03)	
2S071	5	2,802	344	2,458	.88	(−0.97, 2.74)	.69	(−0.99, 2.38)	.19	(−0.09, 0.47)	
2W151	6	5,598	165	5,433	−1.32	(−1.91, −0.73)	−.75	(−1.63, 0.13)	−.57	(−0.93, −0.21)	X
2W171	5	2,681	235	2,446	−.32	(−1.34, 0.70)	−.34	(−1.36, 0.69)	.02	(−0.08, 0.13)	
3E771	1	217	20	197	−1.89	(—, —)	−1.33	(—, —)	−.56	(—, —)	
3P051	6	26,460	695	25,765	−1.27	(−2.03, −0.51)	−1.20	(−1.93, −0.47)	−.07	(−0.17, 0.02)	
3P071	5	5,820	522	5,298	−1.33	(−2.46, −0.21)	−1.32	(−2.34, −0.31)	−.01	(−0.14, 0.13)	
4N051	6	5,424	137	5,287	−.06	(−1.61, 1.49)	−.01	(−1.37, 1.35)	−.05	(−0.29, 0.18)	
4N071	5	1,945	240	1,705	−1.25	(−2.22, −0.27)	−1.20	(−2.24, −0.17)	−.04	(−0.17, 0.09)	

Note. Test = Label for test or test category included in the meta-analysis; *k* = Number of test administrations that contributed to the meta-analysis; *N* = Total sample size; n_{Focal} = Size of focal-group examinee sample; $n_{Referent}$ = Size of referent-group examinee sample; Mean = Random-effects meta-analytic average path coefficient; 95% CI = 95% confidence interval around the meta-analytic mean, Sig. Indirect = Indicates with an "X" which indirect effects had confidence intervals that excluded zero. Negative values indicate that the referent group's mean was higher than the focal group's mean.