AFCAPS-TR-2020-0001

Identify Potential I/O or Non-I/O Psychology Assessment Tools/Methods: AFOQT Methods to Reduce Adverse Impact

USAF Strategic Personnel Research Program

May 2020

C. Wayne Shore, Ph.D.; Natasha Haight, Luisa Martinez

Air Force Personnel Center
Strategic Research and Assessment
HQ AFPC/DSYX

Prepared for:

Katie Gunther, Ph.D.

Air Force Personnel Center
Strategic Research and Assessment Branch

Air Force Personnel Center
Strategic Research and Assessment
HQ AFPC/DSYX
550 C Street West, Ste 45
Randolph AFB TX 78150-4747

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report was cleared for release by HQ AFPC/DSYX Strategic Research and Assessment Branch (SRAB) and is releasable to the Defense Technical Information Center.

This report is published as received with minor grammatical corrections. The views expressed are those of the authors and not necessarily those of the United States Government, the United States Department of Defense, or the United States Air Force. In the interest of expediting publication of impartial statistical analysis of Air Force tests SRAB does not edit nor revise Contractor assessments appropriate to the private sector which do not apply within military context.

Federal Government agencies and their contractors registered with Defense Technical Information Center should direct request for copies of this report to:

Defense Technical Information Center - http://www.dtic.mil/

Approved for public release, unlimited distribution by AFPC/DSYX Strategic Research and

Assessment Branch (SRAB) Joint Base San Antonio-Randolph AFB, TX 78150-4747 or higher DoD authority. Please contact

AFPC/DSYX Strategic Research and Assessment Branch (SRB) with any questions or concerns with the report.

This paper has been reviewed by the Air Force Center for Applied Personnel Studies (AFCAPS) and is approved for publication. AFCAPS members include: Senior Editor Dr. Thomas Carretta AFMC 711 HPW/RHCI and Dr. Imelda Aguilar HQ AFPC/DSYX.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
05-01-2020	Technical Report	October 2014 – June 2019

4. TITLE AND SUBTITLE
Identify Potential I/O or Non-I/O Psychology Assessment Tools/Methods: AFOQT Methods to Reduce Adverse Impact

5a. CONTRACT NUMBER
HCats# GS02Q17DCR0008

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
C. Wayne Shore, Ph.D., Natasha Haight, Luisa Martinez

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
HQ Air Force Personnel Center
Strategic Research and Assessment Branch
550 C. Street West, JBSA-Randolph AFB, TX. 78150-4747

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
HQ Air Force Personnel Center
Strategic Research and Assessment Branch
550 C. Street West, JBSA-Randolph AFB, TX. 78150-4747

10. SPONSOR/MONITOR'S ACRONYM(S)
HQ AFPC/DSYX

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFCAPS-TR-2020-0005

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release. Distribution is Unlimited.

13. SUPPLEMENTARY NOTES

The Air Force Officer Qualifying Test (AFOQT) benefits both the Air Force, by increasing the efficient use of its human resources, and also the examinees, by giving them an opportunity to demonstrate their capability for serving as Air Force officers. Both parties benefit from and have a vested interest in a valid testing process. These are the positive impacts of the AFOQT. A natural outcome of tests is that subgroups perform similarly but not identically. Even when the scores of subgroups mostly overlap, sometimes the mean differences are substantial enough to be of concern when members of a subgroup are less often considered qualified. The test is then said to have adverse impact for the subgroup with the lower average scores. An issue addressed in this study is whether adverse impact for AFOQT subtests is an artifact of AFOQT content and/or its development. The prevalence of adverse impact for other similar tests provides substantial evidence that that is not the case. Adverse impact is common to all large-scale testing programs. This study explores implementable methods to reduce adverse impact of the AFOQT. This study reviewed academic research of methods to mitigate adverse impact with the goal of adopting successful methods for AFOQT subtests.

15. SUBJECT TERMS
Air Force Officer Qualifying Test (AFOQT), Adverse Impact, Subgroup Differences

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
Unclassified					Katie Gunther, Ph.D.
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	**U**	**U**	**U**	125	210-565-5245

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

ACKNOWLEDGMENTS ..... **VII**
FOREWORD ..... **VIII**
EXECUTIVE SUMMARY ..... **IX**
1.0 INTRODUCTION **1**
1.1 Issue 1
1.2 Purpose of this Study 1
2.0 SUBGROUP DIFFERENCES IN THE LITERATURE AND AFOQT **1**
2.1 Approach to Making Construct Comparisons 1
2.2 Findings 2
2.2.1 Black-White Differences in Mean Scores 3
2.2.2 Hispanic/Non-Hispanic Differences in Mean Scores 5
2.2.3 Asian-White differences in Mean Scores 6
2.2.4 Female-Male Differences in Mean Scores 8
2.2.5 Aviation Information-All Groups 9
3.0 SUBGROUP DIFFERENCE AMELIORATION IN THE LITERATURE **10**
3.1 AFOQT Tests Showing Adverse Impact 10
3.2 Candidate Methods for Subgroup Difference Reduction 11
3.2.1 Promising Candidate Methods 11
3.2.2 Other Candidate Methods for Subgroup Difference Reduction 14
3.2.3 Candidate Methods Not Recommended 16
4.0 DISCUSSION **21**
4.1 The AFOQT's Positive Impact 21
4.2 Goal: Reducing Adverse Impact 21
4.3 Prevalence of Adverse Impact 22
4.4 Important Considerations for Evaluating a Selection and Classification Process 22
4.5 Adverse Impact and validity – Two Scenarios 22
4.6 Evaluating the AFOQT at Total Examinee vs. Subgroup Levels 23
4.7 Validity Considerations 24
4.8 Steps Required to Track Validity with Test Changes 24
4.9 Adverse Impact at the Item, Test, and Composite Level 25
4.10 False Flag? 25
4.11 Levels of Effort Required 26

4.12 Possible Actions, Likely Outcomes, and Comments 26
4.13 Finally 28
5.0 CONCLUSIONS **29**
1.0 EXTENDED INTRODUCTION CONTENT **32**
1.1 Background 32
1.1.1 AFOQT 32
1.1.2 Adverse Impact 36
1.2 Method and Approach 37
2.0 EXTENDED DISCUSSION ON SUBGROUP DIFFERENCES IN THE LITERATURE AND AFOQT **37**
2.1 Approach to Making Construct Comparisons 37
2.2 Comparison Considerations 38
2.3 Findings 39
2.3.1 Black-White Differences in Mean Scores 41
2.3.2 Hispanic/Non-Hispanic Differences in Mean Scores 46
2.3.3 Asian-White differences in Mean Scores 49
2.3.4 Female-Male Differences in Mean Scores 53
2.3.5 Aviation Information - All Groups 59
3.0 EXTENDED DISCUSSION OF METHODS TO REDUCE ADVERSE IMPACT **60**
3.1 Subtests for which Adverse Impact Exists 60
3.2 Subgroup Difference Amelioration Methods 61
3.2.1 Verbal Load 61
3.2.2 Stereotype Threat 65
3.2.3 Adding Low-Impact Tests 68
3.2.4 Golden Rule-Type Adjustments 73
3.2.5 Alternative Measures 74
3.2.6 Structured Interviews 76
3.2.7 Impacts of Differing Item Answering Strategies 77
3.2.8 Adjusting for Differential Test Exposure 79
3.2.9 Item and Response Types 81
3.2.10 Response Styles 85
3.2.11 Item Content Changes 86
3.2.12 Constructed Response 88
4.0 OTHER DISCUSSION CONTENT **90**
4.1 Limitations and Future Research 90

4.2 Reduction Methods Outside Our Scope 91
FURTHER READING **93**
REFERENCES **95**

ACKNOWLEDGMENTS

We offer our sincere thanks to the following individuals for their valuable efforts toward the completion of this project.

Daniela Peña (Project Assistant, Editor)

Darien Wolliston (Contributor)

Benjamin Fairbank, Ph.D. (Editor)

Malcolm Ree, Ph.D. (Editor)

Lori Knutson (Editor; ACLC)

Barbara Shore (Technical Assistant)

FOREWORD

There are two areas addressed in this report, a literature review of research concerning methods for reducing adverse impact, and recommendations for implementing those methods that could potentially reduce adverse impact for the AFOQT. The literature review is straightforward, and the methods found are cited and discussed briefly in Section 3.0. The information from that review is reported in detail in the Supplementary Information section.

There are two major dimensions concerning recommendations. First, there are methods that appear to merit priority consideration for implementation. Second, there are many important policy issues to consider associated with a strategic plan to reduce adverse impact in the Air Force. We believed it was our responsibility to discuss those issues while identifying methods to consider. Those considerations are covered most fully in the Discussion section, and briefly in the Executive Summary and the Conclusions.

EXECUTIVE SUMMARY

The Air Force Officer Qualifying Test (AFOQT) benefits both the Air Force, by increasing the efficient use of its human resources, and also the examinees, by giving them an opportunity to demonstrate their capability for serving as Air Force officers. Both parties benefit from and have a vested interest in a valid testing process. These are the positive impacts of the AFOQT.

A natural outcome of tests is that subgroups perform similarly but not identically. Even when the scores of subgroups mostly overlap, sometimes the mean differences are substantial enough to be of concern when members of a subgroup are less often considered qualified. The test is then said to have adverse impact for the subgroup with the lower average scores. An issue addressed in this study is whether adverse impact for AFOQT subtests is an artifact of AFOQT content and/or its development. The prevalence of adverse impact for other similar tests provides substantial evidence that that is not the case. Adverse impact is common to all large-scale testing programs.

The AFOQT has a major effect on the direction of many lives, requiring that adverse impact be taken very seriously to determine that the test battery is justified by its predictive accuracy and its resulting benefits.

Adverse impact occurs frequently, but it is an undesirable condition that could reflect a possible inequitable attribute of a test. Personnel test managers have a responsibility to ensure that test results are not even partially a function of subgroup characteristics unrelated to the construct being tested. Test bias is a potential source of test inequity, but it is independent of adverse impact and is not addressed here.

This study explores implementable methods to reduce adverse impact of the AFOQT. This study reviewed academic research of methods to mitigate adverse impact with the goal of adopting successful methods for AFOQT subtests. Methods to reduce adverse impact need to be evaluated to determine that they retain or enhance validity and that there are no unintended negative consequences. A detrimental outcome of some adverse impact mitigation practices, such as reducing the extent of testing, would be to reduce the benefits of examinees' merits and reduce their likelihood of occupational success. Lower validity for any subgroup would have an immediate and negative effect on both the Air Force and the examinees.

Some mitigation methods may reduce adverse impact for one subgroup while at the same time increasing it for others. That is a condition to be avoided for both the subtests and their use in a composite score. We considered these outcomes in our evaluation of mitigation methods.

Some methods were identified which may have a small but positive effect in reducing adverse impact, while not being expected to have any negative effects such as reducing validities for any of the subgroups or creating adverse impact for any subgroup. These approaches included expanding some subtest's instructions and adding measures expected to have no adverse impact. An Air Force-wide study and a review of a National Academy of Science study, both initiated by the Air Force's Strategic Personnel Research Program, were recently concluded with recommendations of additional constructs that should be evaluated for operational testing. If adopted and confirmed by research, those recommendations may result in a more valid testing program with less adverse impact.

The AFOQT was developed and implemented for the entire population of examinees as one entity without the data which would show subgroup differences in test performance or validity. The issue of adverse impact and its associated issues creates a substantial expansion of the personnel testing program, including a need for creating policies guiding tradeoffs between conflicting outcomes, such as decreasing both adverse impact and validity. A greater evaluation effort would be required to determine the related effects of mitigation methods.

1.0 INTRODUCTION

1.1 Issue

The Air Force Officer Qualifying Test (AFOQT) is an effective means of determining qualification for service as an officer in the Air Force as well as for identifying those job assignments for which qualified applicants are well suited. Although the selection and classification purpose of the AFOQT is achieved, an unwanted outcome for some subtests is that some minority subgroups, defined by race, ethnicity, and gender, perform markedly less well on average than the majority, even though there is nearly always a large overlap in scores. When the difference reaches a threshold magnitude, that condition is known as adverse impact and is the subject of this study.

1.2 Purpose of this Study

The overall purpose of this study was to identify and recommend testing methods for reducing or mitigating adverse impact for the AFOQT subtests. This study primarily reviews recent research studies dealing with the reduction of adverse impact. The studies were evaluated to determine their possible application to the AFOQT based on these criteria:

- the extent to which adverse impact is reduced for the group in question,
- whether the mitigation method increases adverse impact for any group,
- the impact of a recommended method on test validity, and
- whether the level of effort required for implementation for any of the AFQOT subtests is feasible.

2.0 SUBGROUP DIFFERENCES IN THE LITERATURE AND AFOQT

2.1 Approach to Making Construct Comparisons

The occurrence of adverse impact in AFOQT subtests was compared to the occurrence of adverse impact in other similar major tests, such as the GRE, SAT, and ACT. Comparisons between effect size estimates in AFOQT subtests and estimates found in the literature were made whenever sufficiently similar constructs were found. Recent large scale or meta-analytic estimates were prioritized. Assessments for verbal and quantitative ability frequently reported data only for

composites of multiple specific measures. In such cases, comparisons were made with the AFOQT Form T Verbal and Quantitative operational composites instead of a specific subtest. Block Counting had no approximate matches, so spatial tasks were used as the best available comparisons, with notes on differing content. Instrument Comprehension was also compared to spatial tests, because of the test's similarities to spatial rotation and spatial orientation tests. Sufficiently similar tests were found for physical science and perceptual speed allowing for direct comparisons. The SDI-O subtests were compared to findings in the Big Five personality dimensions, as well as any Big Five facets or other similar inventories for which findings were available.

2.2 Findings

Table 1 presents a representative summary of data found outside the Air Force regarding the achievement constructs measured in the AFOQT. The Air Force defines the threshold of concern for adverse impact as when subgroup differences reach $d = 0.40$. Each X in the table denotes the presence of adverse impact.

Tables that include the exact effect sizes in Table 1 can be found in the Supplementary Information section 2.3. Other detailed findings are presented in the Supplementary Information sections 2.3.1 to 2.3.5.

The sections that follow Table 1 compare effect sizes from the AFOQT to other effect sizes by subgroup, still based on the adverse impact threshold. Note that comparisons in these sections were made only when data were available for the subtest, so not every table includes every AFOQT subtest. The SDI-O measures were not included in this table.

Table 1. Summary of Presence/Absence of Adverse Impact in the AFOQT Constructs

Subtest	AFOQT				SAT				GRE				ACT			**	NAEP				Meta-analysis				MCAT			**	FAA selection			
Subgroup*>	B	H	A	F	B	H	A	F	B	H	A	F	B	H	F	A	B	H	A	F	B	H	A	F	B	H	F	A	B	H	A	F
Verbal Composite	X	X	X	-					X	X	-	-									X	X		-								
Verbal Analogies	X	-	-	-																				-					X	X	-	-
Reading Comprehension	X	-	X	-	X	X	-	-					X	X			X	X	-	-				-								
Word Knowledge	X	-	X	-													X	X	-	-				-								
Quantitative Composite	X	-	-	X	X	X	**O**	-	X	X	**O**	X	X	X			X	X	-	-	X	-		-								
Arithmetic Reasoning	X	-	-	X																									X	X	-	X
Math Knowledge	X	-	-	-																												
Table Reading	X	-	-	-																				**O**					X	-	-	-
Block Counting	X	-	-	X																	X			X								
Instrument Comprehension	X	-	X	X																	X			X								
Physical Science	X	-	-	X									X	X		-	X	X	-	-					X	X		-				
Aviation Information	X	-	X	X																												

Note: X = AI against minority group/females, O = AI against majority group/males. SAT = Administration in 2016 (College Board, 2016); GRE = All examinees from July 2017 to June 2018; ACT = all examinees from 1997-2005 & 2007; NAEP = National Assessment of Educational Progress, a national standardized examination given to a national probability sample of students in the U.S--Twelfth grade data from 2015 used here; Meta-analysis = all race data from Roth et al. (2001) except Block Counting and Instrument Comprehension Comparisons for Black examinees from Schmitt et al. (1996), Female data for verbal constructs from Hyde & Linn, (1988), female data for quantitative constructs from Else-Quest et al. (2010), female data for spatial rotation (compared to Block Counting and Instrument Comprehension) from Maeda & Yoon (2013), perceptual speed (compared to Table Reading) from Hedges & Nowell, (1995); MCAT = Administration in 2009 (Davis et al., 2013); FAA selection = Barrier analysis conducted on the Federal Aviation Administration's selection processes for Air Traffic Control Specialist applicants (Outtz & Hanges, 2013).

* B: Black; H: Hispanic; A: Asian; F: Female

** ACT and MCAT administered in 1998 (Camara & Schmidt, 1999)

2.2.1 Black-White Differences in Mean Scores

Table 2 provides summary information on the standardized mean differences found between Black and White individuals across the literature reviewed, compared to the values found in the AFOQT. The left side of the table includes the standardized differences for the main constructs and the right side includes any specific subtests within the construct to the left that were available in the

literature. The highlighted rows show the Cohen's *d* values found in the literature, with superscript marking the source in the notes. All effect size estimates in the AFOQT and those from other measures showed adverse impact against Black examinees.

Table 2. Black-White Standardized Differences in the Literature and AFOQT: Aptitude Assessments

Main Construct			Specific Subtests			
			Verbal Analogies	**Reading Comprehension**	**Word Knowledge**	**Arithmetic Reasoning**
Verbal Composite	AFOQT	**0.95**	**0.86**	**0.97**	**0.77**	
	Literature	**0.94**[b], **0.83**[d]	**0.83**[g]	**0.95**[a], **0.86**[c], **0.75**[h]	**0.82**[h]	
Quantitative Composite	AFOQT	**0.85**				**0.93**
	Literature	**1.04**[a], **0.95**[b], **0.90**[c], **0.74**[d], **0.95**[h]				**1.13**[g]
Physical Science	AFOQT	**0.87**				
	Literature	**0.80**[e], **0.97**[c], **1.04**[h]				
Spatial Tests	AFOQT	**1.03** (BC), **1.15** (IC)				
	Literature	**0.66**[f]				
Perceptual Speed	AFOQT	**0.82**				
	Literature	**0.47**[g]				

Note: Positive values indicate that White examinees scored higher, and negative values indicate that Black examinees scored higher. Effect sizes meeting the $d = |0.40|$ threshold are in bold. BC = Block Counting, IC = Instrument Comprehension. For the AFOQT, White $N = 25{,}148$, Black $N = 3{,}308$.

[a]The class of 2016 who took the SAT. Data included for the critical reading and mathematics sections. White $N = 742{,}436$, Black $N = 199{,}306$ (College Board, 2016).

[b]U.S. citizens who took the GRE from July 2017-June 2018 using most recent test scores. Data included for the verbal reasoning and quantitative reasoning sections. Black $N = 26{,}665$, White $N = 182{,}623$ (Educational Testing Service, 2018).

[c]All White and Black examinees for the ACT over 1997-2005, and 2007 (Sackett & Shen, 2010).

[d]Meta-analytic estimate from applicant industrial samples (Roth et al., 2001).

[e]Examinees who took the MCAT in 2009 using their most recent test scores. White $N = 33{,}807$, Black $N = 6{,}183$ (Davis et al., 2013).

[f]Meta-analytic estimate of spatial ability (Schmitt et al., 1996).

[g]Sample of applicants taking tests administered by the Federal Aviation Administration for Air Traffic Control Specialist selection from 2006-2011. White $N = 10{,}035$, Black $N = 3{,}197$ (Outtz & Hanges, 2013).

[h]NAEP National data for 12th-grade students from the year 2015 for physical science, the mathematics composite, the reading comprehension composite, and the meaning vocabulary scale.

Comparisons for the personality measures were found only for Stress Under Pressure and Dominance-Leader (see Foldes et al., 2008). None of the personality measures from the AFOQT or other measures had differences that met the threshold of $d = |0.40|$.

Caution should be taken when interpreting these and other personality comparisons throughout Section 2.0 as relationships with performance may be curvilinear. The most desirable level of a trait is not necessarily known.

2.2.2 Hispanic/Non-Hispanic Differences in Mean Scores

Table 3 provides summary information on the standardized mean differences found between Hispanic and White or Hispanic and non-Hispanic individuals across the literature reviewed, compared to the values found in the AFOQT. The left side of the table includes the standardized differences for the main constructs and the right side includes any specific subtests within the construct to the left that were available in the literature. The highlighted rows show the Cohen's *d* values found in the literature, with superscript marking the source in the notes.

Almost all verbal and quantitative effect size estimates found in the literature showed adverse impact against Hispanic test takers, but only the composite of the verbal subtests in the AFOQT reached the threshold for adverse impact. The same pattern held for physical science estimates, where only the AFOQT estimate did not reach the adverse impact threshold. No adverse impact was found against Hispanic examinees for perceptual speed in the AFOQT (measured with Table Reading) or a perceptual speed measure administered by the Federal Aviation Administration. Note that for the external measures, effect size was calculated between White and Hispanic means, while in the AFOQT, effect size was calculated between Hispanic and non-Hispanic means.

As with the Black-White comparisons, only the personality measures Stress Under Pressure and Dominance-Leader had appropriate comparisons with other measures (see Foldes et al., 2008). None of the personality measures had differences that met the threshold of $d = |0.40|$. The external estimates found in the literature were calculated between White and Hispanic means, while the AFOQT estimates were calculated between non-Hispanic and Hispanic means.

Table 3. Hispanic/non-Hispanic Standardized Differences in the Literature and AFOQT: Aptitude Assessments

Main Construct			Specific Subtests			
			Verbal Analogies	Reading Comprehension	Word Knowledge	Arithmetic Reasoning
Verbal Composite	AFOQT	**0.40**	0.39	0.33	0.35	
	Literature	**0.56**[b], **0.40**[d]	**0.74**[f]	**0.77**[a], **0.55**[c], **0.49**[g]	**0.63**[g]	
Quantitative Composite	AFOQT	0.35				0.35
	Literature	**0.77**[a], **0.51**[bc], 0.28[d], **0.66**[g]				**0.76**[f]
Physical Science	AFOQT	0.28				
	Literature	**0.60**[e], **0.61**[c], **0.72**[g]				
Spatial Tests	AFOQT	0.22 (BC), 0.28 (IC)				
	Literature	-				
Perceptual Speed	AFOQT	0.26				
	Literature	0.28[f]				

Note: Positive values indicate that non-Hispanic/White examinees scored higher, and negative values indicate that Hispanic examinees scored higher. Effect sizes meeting the $d = |0.40|$ threshold are in bold. For the AFOQT, non-Hispanic $N = 33{,}580$, Hispanic $N = 5731$.

[a]The class of 2016 who took the SAT. Data included for the critical reading and mathematics sections. White $N = 742{,}436$, Hispanic $N = 355{,}829$ (College Board, 2016).

[b]U.S. citizens who took the GRE from July 2017-June 2018 using their most recent test scores. Data included for the verbal reasoning and quantitative reasoning sections. Hispanic $N = 30{,}539$, White $N = 182{,}623$ (Educational Testing Service, 2018).

[c]All White and Hispanic examinees for the ACT over 1997-2005, and 2007(Sackett & Shen, 2010).

[d]Meta-analytic estimate from industrial samples. $N = 6{,}133$ (Roth et al., 2001).

[e]Examinees who took the MCAT in 2009 using their most recent test scores. White $N = 33{,}807$, Hispanic $N = 5{,}810$ (Davis et al., 2013).

[f]Sample of applicants taking tests administered by the Federal Aviation Administration for Air Traffic Control Specialist selection from 2006-2011. White $N = 10{,}035$, Hispanic $N = 940$ (Outtz & Hanges, 2013).

[g]NAEP National data for 12th-grade students from the year 2015 for physical science, the mathematics composite, the reading comprehension composite, and the meaning vocabulary test.

2.2.3 Asian-White differences in Mean Scores

Table 4 provides summary information on the standardized mean differences found between Asian and White individuals across the literature reviewed, compared to the values found in the AFOQT. The left side of the table includes the standardized differences for the main constructs and the right side includes any specific subtests within the construct to the left that were available in the literature. The highlighted rows show the Cohen's *d* values found in the literature, with superscript marking the source in the notes.

Two of three verbal subtests in the AFOQT and the composite of these subtests had adverse impact against Asian examinees, while no other measures did. None of the quantitative ability measures

or physical science measures had adverse impact against Asian examinees. Instrument Comprehension in the AFOQT had adverse impact against Asian examinees, but Block Counting in the AFOQT and two out of three spatial ability measures found in the literature did not. The perceptual speed measure in the AFOQT (i.e., Table Reading) did not have adverse impact against Asian examinees, and one of the three external perceptual speed measures did.

Table 4. Asian-White Standardized Differences in the Literature and AFOQT: Aptitude Assessments

Main Construct			Specific Subtests			
			Verbal Analogies	Reading Comprehension	Word Knowledge	Arithmetic Reasoning
Verbal Composite	AFOQT	**0.49**	0.34	**0.54**	**0.40**	
	Literature	0.10[b]	0.22[j]	-0.01[a], -0.05[k]	0.13[k]	
Quantitative Composite	AFOQT	-0.26				-0.10
	Literature	**-0.63**[a], **-0.41**[b], -0.29[k]				0.13[j]
Physical Science	AFOQT	0.11				
	Literature	0.29[c], 0.04[d], -0.23[k]				
Spatial Tests	AFOQT	0.15 (BC), **0.40** (IC)				
	Literature	**0.79**[e], 0.34[f], 0.14[g]				
Perceptual Speed	AFOQT	0.10				
	Literature	0.12[h], **0.82**[i], -0.03[j]				

Note: Positive values indicate that White examinees scored higher, and negative values indicate that Asian examinees scored higher. Effect sizes meeting the $d = |0.40|$ threshold are in bold. BC = Block Counting, IC = Instrument Comprehension. For the AFOQT, White $N = 25{,}148$, Asian $N = 1{,}894$.

[a]The class of 2016 who took the SAT. Data included for the critical reading and mathematics sections. White $N = 742{,}436$, Asian $N = 196{,}735$ (College Board, 2016).

[b]U.S. citizens who took the GRE from July 2017-June 2018 using their most recent test scores. Data included for the verbal reasoning and quantitative reasoning sections. Asian $N = 22{,}567$, White $N = 182{,}623$ (Educational Testing Service, 2018).

[c]individuals who took the MCAT Physical Sciences test in 1998 (Camara & Schmidt, 1999).

[d]individuals who took the ACT Science test in 1998 (Camara & Schmidt, 1999).

[e]Sample of young adults in the U.S. (mean age = 20.1) and China (mean age = 19.4) who took the Card Rotations test. Chinese $N = 40$, American $N = 66$ (Geary et al., 1996).

[f]Sample of young adults in the U.S. (mean age = 20.1) and China (mean age = 19.4) who took the Cube Rotations test. Chinese $N = 40$, American $N = 66$ (Geary et al., 1996).

[g]Sample of young adults in the U.S. (mean age = 20.1) and China (mean age = 19.4) who took the Mental Rotations test. Chinese $N = 40$, American $N = 66$ (Geary et al., 1996).

[h]Sample of young adults in the U.S. (mean age = 20.1) and China (mean age = 19.4) who took the Number Comparisons test. Chinese $N = 40$, American $N = 66$ (Geary et al., 1996).

[i]Sample of young adults in the U.S. (mean age = 20.1) and China (mean age = 19.4) who took the Identical Pictures test. Chinese $N = 40$, American $N = 66$ (Geary et al., 1996).

[j]Sample of applicants taking tests administered by the Federal Aviation Administration for Air Traffic Control Specialist selection from 2006-2011. White $N = 10{,}035$, Asian $N = 517$ (Outtz & Hanges, 2013).

[k]NAEP National data for 12th-grade students from the year 2015 for physical science, the mathematics composite, the reading comprehension composite, and the meaning vocabulary scale.

As with the other minority groups, comparisons for the personality measures were found only for Stress Under Pressure and Dominance-Leader. Unlike the other minority groups, the AFOQT measure Stress Under Pressure met the $d = |0.40|$ threshold. None of the other measures met the threshold, although one of two comparison measures found for Stress Under Pressure, Even Tempered, was just below the threshold and in the same direction as Stress Under Pressure (Foldes et al., 2008).

2.2.4 Female-Male Differences in Mean Scores

Table 5 provides summary information on the standardized mean differences found between female and male individuals across the literature reviewed, compared to the values found in the AFOQT. The left side of the table includes the standardized differences for the main constructs and the right side includes any specific subtests within the construct to the left that were available in the literature. The highlighted rows show the Cohen's *d* values found in the literature, with superscript marking the source in the notes.

No verbal ability measures from the AFOQT or other measures had adverse impact against women. Arithmetic Reasoning and the Quantitative composite (which includes Arithmetic Reasoning) in the AFOQT showed adverse impact against women, and two of six external measures did. The Physical Science test in the AFOQT had adverse impact against women, but an external measure did not. All spatial measures showed adverse impact against women. No perceptual speed measures showed adverse impact against women.

More personality measure comparisons were able to be made for female-male differences than for the minority group differences. Appropriate comparisons were found for Stress Under Pressure, Dominance-Leader, Unassertive, Hyper-Competitive, and Team Player. Overall, effect sizes for personality traits did not pass $d = |0.40|$ in the AFOQT or other measures. However, Stress Under Pressure and its comparison from the literature both met the threshold in the same direction. Hyper-competitive in the literature passed the threshold in one study found (at $d = 0.46$; Thornton et al., 2011) but did not in two estimates from another study (Bhardwaj et al., 2018). See Supplementary Information section 3.2.4 under D for specific data and sources for each of the personality comparisons.

Table 5. Female-Male Standardized Differences the Literature and AFOQT: Aptitude Assessments

Main Construct			Specific Subtests			
			Verbal Analogies	**Reading Comprehension**	**Word Knowledge**	**Arithmetic Reasoning**
Verbal Composite	AFOQT	0.29	0.19	0.37	0.24	
	Literature	0.36[b], -0.11[c]	0.16[c], 0.20[i]	0.02[a], -0.03[c], -0.25[f]	-0.02[c], 0.00[f]	
Quantitative Composite	AFOQT	**0.46**				**0.54**
	Literature	0.25[a], **0.59**[b], 0.06[d], 0.15[e], 0.09[f]				**0.72**[i]
Physical Science	AFOQT	**0.66**				
	Literature	0.22[f]				
Spatial Tests	AFOQT	**0.49** (BC), **1.08** (IC)				
	Literature	**0.57**[g]				
Perceptual Speed	AFOQT	0.15				
	Literature	-0.21 or **-0.43**[h], 0.06[i]				

Note: Positive values indicate that male examinees scored higher, and negative values indicate that female examinees scored higher. Effect sizes meeting the $d = |0.40|$ threshold are in bold. BC = Block Counting, IC = Instrument Comprehension. For the AFOQT, male $N = 29{,}536$, female $N = 10{,}550$.

[a]The class of 2016 who took the SAT. Data included for the critical reading and mathematics sections. Male $N = 762{,}247$, female $N = 875{,}342$ (College Board, 2016).

[b]U.S. citizens who took the GRE from July 2017-June 2018 using their most recent test scores. Data included for the verbal reasoning and quantitative reasoning sections. Male $N = 113{,}925$, female $N = 199{,}698$ (Educational Testing Service, 2018).

[c]Meta-analytic review of verbal abilities of individuals from under 4 years old to 26 years and older in the U.S. and Canada for overall verbal ability, vocabulary, and reading comprehension. Total $N = 1{,}418{,}899$ (Hyde & Linn, 1988).

[d]Effect sizes for math performance of grade 11 students calculated from records required under the No Child Left Behind legislation from 10 states, where the mean performance for these states was found to match the national mean. Total $N = 446{,}381$ (Hyde et al., 2008).

[e]Meta-analytic summary of mathematic performance for students aged ~14-16 years across 69 countries (Else-Quest et al., 2010).

[f]NAEP National data for 12th-grade students from the year 2015 for physical science, the mathematics composite, the reading comprehension composite, and the meaning vocabulary scale.

[g]Meta-analytic review of gender differences in the Purdue Spatial Visualization Tests: Visualization of Rotations (Maeda & Yoon, 2013).

[h]Meta-analytic review of 4 large scale probability samples of adolescents and young adults the U.S. Total $N = 127{,}268$ (Hedges & Nowell, 1995).

[i]Applicants taking tests administered by the Federal Aviation Administration for Air Traffic Control Specialist selection from 2006-2011. Male $N = 11{,}813$, female $N = 3{,}635$ (Outtz & Hanges, 2013).

2.2.5 Aviation Information-All Groups

Information on racial, ethnic, and gender differences for Aviation Information was not available in the literature. Data on civilian participation in aviation indicates that racial/ethnic minorities and women lag behind the majority in participation in aviation (Ison, Herron, & Weiland, 2016). This indicates that these groups may not have had as much exposure to aviation concepts overall as the majority groups, which would put them at a disadvantage on tests with aviation content. In the

AFOQT, adverse impact occurred in Aviation Information and Instrument Comprehension for women and the racial minority groups, but not Hispanic examinees.

3.0 SUBGROUP DIFFERENCE AMELIORATION IN THE LITERATURE

3.1 AFOQT Tests Showing Adverse Impact

Table 6 shows which groups had adverse impact ($d \geq 0.40$) for each subtest in the AFOQT. These subtests and groups are the focus of the adverse impact amelioration methods researched and reported below. The full table with numbers is included in section 3.1 of the Supplementary Information.

Table 6. Standardized Mean Differences in AFOQT Subtests for Minority Groups

Subtest	Black-White	Female-Male	Asian-White	Hispanic-Non
Verbal Analogies	✓	-	-	-
Arithmetic Reasoning	✓	✓	-	-
Word Knowledge	✓	-	✓	-
Math Knowledge	✓	-	-	-
Reading Comprehension	✓	-	✓	-
Physical Science	✓	✓	-	-
Table Reading	✓	-	-	-
Instrument Comprehension	✓	✓	✓	-
Block Counting	✓	✓	-	-
Aviation Information	✓	✓	✓	-
Team Player (+)	-	-	-	-
Stress Under Pressure (-)	-	✓	✓	-
Unassertive (-)	-	-	-	-
Hyper-Competitive (-)	-	-	-	-
Dominance-Leader (+)	-	-	-	-

Note: Check marks indicate adverse impact ($d \geq 0.40$) against that subgroup. Personality traits, starting with Team Player, are either coded positively or negatively, depending on whether they are considered desirable traits.

3.2 Candidate Methods for Subgroup Difference Reduction

There are several considerations associated with each adverse impact amelioration method identified, such as the level of effort required to implement, the expected impact on validities, and the extent to which adverse impact would be reduced. The methods most prominently presented for adoption or further consideration were identified based on their low level of required effort to implement and their expected lack of negative impacts on validity. These methods are generally ordered first by level of effort required to implement and second by effect on validities. No single method is expected to substantially reduce adverse impact. Relevant research for these methods is cited in the Supplementary Information section 3.2, and key citations are included here.

3.2.1 Promising Candidate Methods

Modified Instructions. Extensive research from different fields suggests that lowering the verbal reasoning and language requirements of tests lowers their adverse impact against minority races and ethnicities, likely without decreasing validity (e.g., Abedi & Lord, 2001; Christian et al., 2010; Naglieri, 2005; Naglieri & Ford, 2003). The subgroups most affected seem to be individuals who speak English as a second language. The construct measured by the test may change when verbal requirements are reduced, and most methods of putting such a change in place lack substantial direct research. Some proposed methods include simplifying the language used, giving the test in video format, modifying instructions, or using picture-based or picture-supplemented tests. The most promising method appears to be modifying instructions.

Research suggests that difficulty understanding instructions for unfamiliar tests, such as Block Counting, Instrument Comprehension, and Table Reading, can artifactually reduce scores on the exam because examinees adopt inappropriate strategies for solving the problems (see Lohman & Gambrell, 2012). Examinees may not lack the ability, but rather an understanding of a proper approach to the test. Thus, ensuring that test instructions are comprehensive without complex language and including more examples may prove to be effective. Although this method lacks direct research, its ease and general benefit to examinees make it a promising approach.

Removing Pretest Inquiry of Demographics: Stereotype Threat. Anxiety caused by stereotypes regarding one's group (i.e., stereotype threat) has long been seen as a contributor to subgroup differences. Results of research on the effects of mitigating such anxiety are mixed, so no clear conclusions can be made as to its effectiveness. The simplest method of decreasing stereotype threat is by collecting demographic information after the test so that examinees' group identities are not salient while they take the test. However, this method may also produce the smallest effects. Reducing stereotype threat is unlikely to impact validity or increase adverse impact for other groups. Despite the ambiguity in the research and potentially small influence over adverse impact, the relative ease of implementing this method makes it worth consideration.

Add Motivational and Alternative Constructs. Evidence exists suggesting that certain motivational constructs may decrease subgroup differences and increase validity (see Foldes et al., 2008; Teachout et al., 2019; Mattern et al., 2017). Adding motivational constructs may particularly benefit women. Evidence also suggests that measures of thinking ability, such as logic and learning, are valid predictors of performance with low adverse impact (see De Soete et al., 2013). Research does not indicate whether 'thinking' tests are likely to provide incremental validity over the AFOQT.

Two recent studies initiated under the Strategic Personnel Research Program identified constructs that officer (and enlisted) trainers considered important for Air Force training and occupational success. One study (Shore et al., 2019) included an Air Force-wide survey of technical training instructors and field training instructors who both responded to survey questions and often made constructive comments. Several of the attributes and competencies rated as important for officers' success overlap with motivational and 'thinking' constructs that research outside the Air Force have indicated may reduce subgroup differences. These constructs are Achievement and Responsibility in the TAPAS, Reasoning (soon to be available in the Manpower Test Battery), Self-Discipline (available in the SDI-O), and Problem Solving. Another study (Teachout et al., 2019) reviewed the findings of the National Academy of Science which presented suggestions from senior psychologists concerning constructs that might be tested by the military and used for selection and classification. This review also showed overlap with outside research on motivational and 'thinking' measures, with such constructs as Working Memory, Fluid Intelligence, Learning Agility, Achievement, and Responsibility.

We recommend that further consideration be given to evaluating the validity of these tests. Both studies under the Strategic Personnel Research Program also identified other high-priority constructs to increase validity of tests such as the AFOQT, which may act to reduce subgroup differences as well. Accepts Feedback, Active Listening, Professionalism, Adaptability, and Timeliness were some other competencies rated highly by trainers, and Attention to Detail, Adjustment, and Situational Awareness were some attributes rated highly by trainers (Shore et al., 2019). All attributes listed here currently have tests available in the Department of Defense. Note that the competencies varied in how lacking trainers perceived recruits to be in them. We anticipate that some resulting tests may show a useful relationship to training success, with the caution that for those constructs with a more subtle long-term effect, relevant criteria may not be readily available. As these tests measure constructs that are unlike measures usually used in a testing environment, they may not show predictive validity for the usual criteria. Instead, they may be predictive of success in interpersonal relationships, retention in terms of service, or creativity, for example.

Golden Rule Method. While controversial, evidence suggests that examining the size of subgroup differences in items being developed can reduce subgroup differences without harming the content validity of the test (see Kiddler & Rosner, 2002). Such approaches have been deemed Golden Rule-type methods. Evidence was not available regarding criterion-related validity. Despite this shortcoming, the method is still promising.

Table 7 summarizes the above adverse impact amelioration methods. The table columns list the amelioration method, and then several important considerations. After the amelioration method, the next column describes whether implementing the method is likely to change subgroup difference magnitudes for other groups, and what that change is likely to be. Next, a summary of the method's impact on criterion validity based on the literature findings is reported. Finally, key articles related to the method and the location of extended information on the method in the Supplementary Information are included. Unless noted otherwise, all methods shown in Table 7 apply to the AFOQT as a whole.

Table 7. Candidate Methods More Likely to Reduce Subgroup Differences

Amelioration Method	Effect on AI for Other Subgroups	Effect on Validity	Key Studies	Relevant SI Section
Modified Instructions for TR, BC, & IC	Likely none	Likely none	Lohman & Gambrell, 2012	**3.2.1**
Post-test demographic Inquiry	Likely none	Likely none	Shewach et al., 2019	**3.2.2**
Add motivational measures	Reduce as well	Likely increase	Shore et al., 2019; Teachout et al., 2019	**3.2.3**
Add alternative construct measures	Likely reduce as well	Unknown	Shore et al., 2019; Teachout et al., 2019	**3.2.3**
Add other promising constructs	Unknown	Likely increase	Shore et al., 2019; Teachout et al., 2019	**3.2.3**

Note: TR = Table Reading, BC = Block Counting, IC = Instrument Comprehension

3.2.2 Other Candidate Methods for Subgroup Difference Reduction

The following methods, while still promising for ameliorating adverse impact, have a lower priority for further consideration than those in the previous section. Commonly, this is due to additional research and other efforts required for implementation.

Add STAT Measures. Although research is still in the early stages, available evidence suggests that adding creativity, practical skill, and analytical skill measures from the Sternberg Triarchic Abilities Test (STAT) can lower adverse impact across many subgroups and increase the validity of educational tests (e.g., Sternberg, The Rainbow Project Collaborators, 2006). Available evidence for the STAT considers only validity with educational success as the criterion, so it is unclear whether adding STAT measures to the AFOQT will produce the same results. However, the promise this method offers for both decreasing adverse impact across groups and improving validity makes it a worthy candidate for future consideration.

Add Biodata. A strong body of evidence suggests biodata has good validity and low adverse impact (e.g., Bobko, Roth, & Potosky, 1999; Reilly & Chao, 1982). It also has several complications and drawbacks. DIF analyses and validation on applicant samples are critical to ensure that biodata measures have an operationally low adverse impact. Thus, development of an effective biodata measure may be an extensive effort, but likely to provide payoff when implemented correctly.

Replace Block Counting with Assembling Objects. Military research of the ASVAB suggests that the test Assembling Objects had both validity and low subgroup differences for women and racial/ethnic minorities (Held & Carretta, 2013), but research on the AFQT in the Army found that while the test had incremental validity, subgroup differences were increased for Black and female examinees and decreased for Hispanic examinees (Anderson et al., 2011). As a spatial test, Assembling Objects may serve as a viable replacement to Block Counting. However, because past research has been mixed on whether Assembling Objects increases adverse impact for certain groups, additional research is needed to determine if it merits more consideration.

Structured Interviews. Structured interviews have shown good validity and low subgroup differences (e.g., Campion, et al., 1997; Levashina et al., 2014; Bobko, et al., 1999). Some evidence has shown incremental validity over personality and cognitive measures. However, their time and personnel demand for training interviewers and conducting interviews are disadvantages to take into account. They are also susceptible to cognitive and other biases on part of the interviewers and impression management on part of the interviewee. We suggest a limited trial for key occupations.

Table 8 summarizes the above candidate methods. The organization of the table is similar to Table 7, with the addition of a column indicating the difficulty of implementation. Discussions of each of these methods and relevant citations are included in the Supplementary Information section 3.2. Unless noted otherwise, all methods apply to the AFOQT as a whole.

Table 8. Promising Methods Requiring Additional Research

Amelioration Method	Effect on AI for Other Groups	Effect on Validity	Difficulty to Implement	Key Studies	Relevant SI Section
Add STAT measures	Decreases all groups but Asian	Improves	Difficult	Sternberg, The Rainbow Project Collaborators, 2006	**3.2.3**
Add Biodata	Reduce as well	May improve	Difficult	Bobko et al., 2013; Reilly & Chao, 1982; Schmidt & Hunter, 1998	**3.2.3**
Replace BC with Assembling Objects	Mixed	May Improve	Easy	Held & Carretta, 2013; Anderson et al., 2011	**3.2.5**
Add Structured Interview	Reduce as well	May Improve	Very Difficult	Campion, et al., 1997; Levashina et al., 2014; Bobko, et al., 1999; Huffcut & Roth, 1998	**3.2.6**

3.2.3 Candidate Methods Not Recommended

There are several reasons that methods did not merit consideration. All these methods require some level of additional research, with certain methods requiring a great deal. Some are expected to have little beneficial effect and others may have a negative effect on validity. Even when not presently recommended, it is prudent to be aware that they may at some point deserve a second look, perhaps by being adapted in some way for some form of implementation. Extended discussions of these methods are included in the Supplementary Information section 3.2.

Increase Test Time. Slightly more time allowed per item may decrease female-male differences in performance on Block Counting and Instrument Comprehension. The primary evidence of this approach is a meta-analysis that found the smallest female-male differences in a spatial rotation test at a time limit of 40-60 seconds per item, compared to shorter time limits and no time limit (Maeda & Yoon, 2013). Evidence was not available on what impact this approach has on validity, difficulty, or effects on other groups. Extended discussion of this method is in the Supplementary Information section 3.2.7.

Additional Retesting. Retesting has minimal or negative effects on the size of subgroup differences for minority racial/ethnic groups. However, women may improve more at retest than men, and improvements that allow more minority individuals to achieve qualifying scores can still increase diversity even if subgroup differences remain. Evidence suggests criterion validity is not lowered by retesting (Van Iddekinge et al., 2011; Villado et al., 2016). Implementing retesting would be relatively easy, but improvements after allowing one additional retest for the AFOQT (i.e., three attempts instead of two) may be unlikely. Because of the potential negative impacts on minority racial/ethnic groups coupled with minimal benefits, this method was deemed deficient. Extended discussion of this method is in the Supplementary Information section 3.2.8.

SJT Modifications. During the course of the review, several adjustments that could benefit the Situational Judgment Test (SJT) were also uncovered. Research has demonstrated that the size of subgroup differences found in SJTs varies depending on the SJT's design (e.g., Christian et al., 2010; Lievens et al., 2019; McDaniel, et al., 2007). Video or multimedia SJTs have lower subgroup differences than written SJTs; higher fidelity responses have lower subgroup differences than multiple choice; items inquiring what an examinee *would* do have lower subgroup differences than items inquiring what an examinee *should* do; having examinees rate how likely they would be to enact each behavior option has smaller differences than other response methods; and controlling for response styles produces smaller subgroup differences than allowing examinees to have central or extreme response styles. All of these methods likely impact the underlying construct measured by the SJT and thus impact what the SJT is likely to be valid for. An extended discussion of these methods is available in the Supplementary Information section 3.2.9.

Control Response Style. Research indicates that different subgroups differ in their response styles (i.e., whether they tend to use the center of a Likert-type scale, or the extremes, e.g., Harzing et al., 2012). Researchers developed a method of altering scores on Likert-type scales in SJTs such that these different response styles were controlled (McDaniel et al., 2011). The initial research indicates that subgroup differences are lowered, and validity increased, but additional replications are needed. Thus far, this method has been explored only in SJTs, but it may apply to Stress Under Pressure. However, note that the original study found a reduction in group differences for Black examinees, while the groups passing the adverse impact threshold for Stress Under Pressure were Asian and female examinees, so this approach's effectiveness depends on its generalizability.

Because of these uncertainties, this approach requires additional research. Extended discussion of this method is in the Supplementary Information section 3.2.10.

Adjust Item Content: Quantitative Tests. Research on high-stakes achievement tests has found that quantitative items have different sizes of female-male differences depending on the method that must be used to solve the item (e.g., Gallagher et al., 2000; 2002). Women tend to do better on items that require "by-the-book" response approaches, while men tend to excel at items that require "out-of-the-box" thinking to solve. Quantitative items requiring spatial ability to solve also appear to produce larger gender differences in favor of men. Impacts on validity for this approach are unknown, but it appears possible to produce tests of approximately the same difficulty level. Effects on other subgroups are also unknown. This approach would be most appropriate for Math Knowledge and Arithmetic Reasoning. The considerations for which research is lacking make the benefits of this approach uncertain. Additional research is needed. Extended discussion of this method is in the Supplementary Information section 3.2.11.

Add Integrity Test. Adding an integrity test to the AFOQT is unlikely to be an effective strategy. While extensive evidence indicates that integrity tests have low subgroup differences (e.g., Ones & Viswesvaran, 1998), recent research suggests that validity levels may also be low (e.g., Van Iddekinge et al., 2012), and research in the Air Force indicates that there is little demand for higher integrity in officers than already exists (Shore et al., 2019). These doubts cast on integrity tests need to be resolved before implementation is considered. Extended discussion of this method is in the Supplementary Information section 3.2.3.

Constructed Responses. Constructed responses require examinees to produce a response instead of choosing one as in multiple choice. This approach has mixed evidence that suggests substantial subgroup difference reductions may be possible (e.g., Edwards & Arthur, 2007; Lievens et al., 2019), but increases may also occur (Wilson & Zhang, 1998), and the effect may differ across groups (e.g., Edwards & Arthur, 2007 versus Wilson & Zhang, 1998). The use of constructed responses has many drawbacks, such as increased subjectivity and scoring costs. Extended discussion of this method is in the Supplementary Information section 3.2.12.

Replace Stress Under Pressure with an SJT. Evidence suggests that culture and self-stereotyping that vary across groups may influence responses on self-description inventories (e.g., Cadinu & Galdi, 2012; Cai et al., 2011; Uskul et al., 2010). Thus, replacing Stress Under Pressure with a behavioral index (such as an SJT) may avoid unwanted cultural influences. However, research does not exist beyond defining this problem, so the effectiveness of this approach is speculative. Extended discussion of this method is in the Supplementary Information section 3.2.5.

Image Aids. Extensive research from different fields suggests that lowering the verbal reasoning and language requirements of tests lowers their adverse impact against minority races and ethnicities, potentially without decreasing validity (e.g., Abedi & Lord, 2001; Christian et al., 2010; Naglieri, 2005; Naglieri & Ford, 2003). The subgroups benefitting most seem to consist of individuals for whom English is a second language. The construct measured by the test may change when verbal requirements are lowered, and most methods of putting such a change in place lack substantial direct research. One method includes supplementing items with images or diagrams. Applicable subtests would be Arithmetic Reasoning and Physical Science. However, because of the lack of direct research and research on adult populations, this approach needs research. Extended discussion of this method is in the Supplementary Information section 3.2.1.

Coaching & Training: Spatial Tests. Research has suggested that performance on spatial tests varies depending on the effectiveness of the strategies used to complete the items (e.g., Hirnstein et al., 2009). Strategies used may relate to gender, but not true spatial ability, so training prospective examinees to use the most effective strategy for them may decrease female-male differences. Research has also shown that training on a spatial task can generalize to other spatial tasks, which may indicate that spatial training generalizes to performance (Uttal et al., 2013). Research has shown that men and women improve the same amount when trained in spatial skills (Uttal et al., 2013), but tailored strategy training may be able to narrow the performance gap (Stieff et al., 2013). This approach is relevant for Block Counting and Instrument Comprehension. Implementing it would require research into the strategies used by examinees for these items, the efficiency of each strategy, and the effect on each subgroup of training those strategies. Extended discussion of this method is in the Supplementary Information section 3.2.7 and 3.2.8.

Coaching & Training: All Subtests. All subgroups improve with coaching, and evidence is mixed as to whether majority or minority groups tend to improve more (Sackett et al., 2001). Little evidence was available for impacts on validity, a significant shortcoming. While the improvements experienced by all groups may help more individuals from minority groups achieve qualifying scores, the potential that this method may increase subgroup differences combined with missing information for the impact on validity make this method deficient. Extended discussion of this method is in the Supplementary Information section 3.2.8.

Replace Stress Under Pressure. The current review found that different facets of Neuroticism from the Big Five dimensions inventory vary substantially in their degree subgroup differences and that different groups vary in which facets produce the largest differences (see Foldes et al., 2008). Thus, replacing Stress Under Pressure with one of these Neuroticism facets may decrease adverse impact. However, adverse impact may either increase or decrease depending on the subgroup if another stress measure is used, and it is uncertain how validity would be affected. These risks make this method inappropriate. Extended discussion of this method is in the Supplementary Information section 3.2.5.

Special Interest Items. Research has indicated that individuals perform better on items that have content highly familiar to their group (e.g., Carlton & Harris, 1992; O'Neill & McPeek, 1993). Thus, including items specifically designed to draw on topics familiar to certain subgroups may reduce subgroup differences in Reading Comprehension items in the AFOQT. However, little research exists that directly manipulates test item content to decrease subgroup differences, and the effects of this procedure on validity are largely unknown. Further, AFOQT items are developed with a specific standard of not favoring or disfavoring any subgroup based on differential familiarity with words or concepts. Extended discussion of this method is in the Supplementary Information section 3.2.11.

Non-Verbal Analogies. Extensive research from different fields suggests that lowering the verbal reasoning and language requirements of tests lowers their adverse impact against minority races and ethnicities, potentially without decreasing validity (e.g., Abedi & Lord, 2001; Christian et al., 2010; Naglieri, 2005; Naglieri & Ford, 2003). This method would be especially relevant to examinees for whom English is a second language. However, the construct measured by the test

may change if verbal requirements are lowered, and most methods of putting such a change in place lack substantial direct research. One proposed method that has been used with some success in decreasing subgroup differences when selecting gifted children is giving tests in picture format instead of word format (see Lohman & Gambrell, 2012). The most applicable test for this approach is Verbal Analogies, where verbal comparisons can be replaced with picture comparisons. The extent to which methods used with children can be replicated with adults is uncertain, and this approach would require the test to be essentially rewritten. Because of these points, this method is considered unworthy of further consideration. Extended discussion of this method is in the Supplementary Information section 3.2.1.

4.0 DISCUSSION

4.1 The AFOQT's Positive Impact

Air Force personnel testing benefits both the Air Force and its applicants. The benefit to the Air Force is a more effective and efficient use of human resources. The benefit to the applicants is being assigned to training and to occupations for which they are better suited and more likely to succeed. Both the Air Force and its members are stakeholders in a valid personnel testing process. The Air Force thus has a responsibility to select individuals using the most effective process practical for all the racial, ethnic, and gender subgroups of its members. Modifications of Air Force selection and classification tests should be with the perspective and understanding of the positive impact they provide to the Air Force and to the examinees.

4.2 Goal: Reducing Adverse Impact

Even a highly effective testing process with substantial subgroup differences requires an effort to determine if there are feasible methods for reducing the disparity. This goal has remained elusive in the national testing community. This paper presents some suggestions for further efforts toward that goal.

The adverse impact of the AFOQT and other large-scale tests on examinee subgroups is wide-spread and persistent. In response, much research has been devoted to developing methods to

reduce adverse impact. The primary purpose of this study is to discover current research-based methods for reducing or mitigating adverse impact. An evaluation of these methods requires consideration of related factors which will be discussed here.

4.3 Prevalence of Adverse Impact

An initial question about adverse impact resulting from the AFOQT and its subtests is whether it is a function unique to the AFOQT or whether similar tests have shown similar results. They mainly do. Measures of the same or similar constructs as the ones measured by the AFOQT result in similar levels of adverse impact, with only minor exceptions. The inference is that the AFOQT does not cause subgroup differences, but simply reflects them.

4.4 Important Considerations for Evaluating a Selection and Classification Process

The objective value of a selection and classification process is based on the strength of its relationship with subsequent performance, or its validity. Part of a conceptual validity of overall performance is a measure of a portion of performance. The AFOQT provides that measure. A critical consideration for an Air Force Test is that it must be unbiased against subgroups, and as an objective measure, the AFOQT provides a means for detecting and avoiding bias. However, the AFOQT does not prevent adverse impact, nor does any other demonstrably effective process. Any Air Force process to select personnel other than objective tests must be able to demonstrate validity and a lack of bias against any subgroup.

4.5 Adverse Impact and validity – Two Scenarios

An attempt to mitigate adverse impact for a subgroup through the modification of a test or its use must consider the resulting effect on selection and classification for that subgroup. The critical factor for successful selection and classification is test validity. Research cited in this report has indicated that methods to reduce adverse impact may lessen the validity of the test. To illustrate the importance of retaining validity when implementing a method to reduce adverse impact, consider two hypothetical tests, both with adverse impact present for a given subgroup. For the first test, the validity for the subgroup is very high, which means the test is providing a good benefit for the subgroup as well as for the Air Force. It is placing people where they are likely to succeed

while avoiding training or jobs for them for which they are unlikely to succeed. The first test is performing its function *within* a subgroup, and the concern for adverse impact is *between* the majority and the subgroup. Retention of the *within* subgroup benefit of validity should be a primary goal.

For the second test, there is also adverse impact for the subgroup, but for this test, the validity for that subgroup is very low relative to the validity for the majority group. Therefore, not only is there adverse impact for the subgroup, but the test provides only little or no benefit to the subgroup or to the Air Force. We contend that the lack of validity for a subgroup is detrimental to actions concerning their selection and job placement, whereas adverse impact alone would not be the source of the main concern about the subgroup. Therefore, the main priority of a test should be maximizing validity, and any method of reducing adverse impact that also decreases validity should be avoided. Consequently, validity was a prominent part of the research studies reported here about adverse impact.

4.6 Evaluating the AFOQT at Total Examinee vs. Subgroup Levels

The above examples demonstrate another important point regarding the relative validity of a test for a minority and the majority. In creating a selection and classification testing process, the Air Force has followed the most direct and efficient approach by developing a system that treats the entire population as one entity for which the process is optimized. Personnel tests are implemented because of their demonstrated validity with the *entire population* of examinees. That is appropriate since the operational administration and use of the test results are the same for all examinees regardless of subgroup membership. The presence of adverse impact is not known before tests are implemented and before subgroup test performance data are available. Going beyond having one level of policies and practices uniformly applied to all personnel introduces another dimension that requires attention, namely the issue of potentially differing validity for each subgroup. The question raised by this possibility is whether the test serves the subgroups' interests in terms of validity. That is, what is the validity for each subgroup? The answer to this question is unknown, but not unknowable, as data are available to be analyzed. If the results of the AFOQT are to be treated by examining subgroup performance, this is the most important issue to resolve.

We suggest that adverse impact for a subgroup cannot be fully addressed without considering the associated validity for each subgroup. The condition justifying the highest priority for any remedy would be both the presence of adverse impact for the subgroup and the lack of validity for the subgroup relative to the whole group. We believe that a comprehensive plan to determine subgroup validities would represent a very large effort.

4.7 Validity Considerations

Any action which reduces test effectiveness may have unintended consequences. A less effective personnel selection process denies some more-qualified people the opportunity to demonstrate their ability to succeed and is more likely to deny them the training and assignment that they merit. They will be replaced by those less likely to succeed, a negative event in the lives of all persons who are less well served by the test. Psychometricians have a responsibility to provide effective tests that help guide examinees to more likely success in training and occupations.

A potentially important consideration in this context relates to the following questions: If the test results in adverse impact for a subgroup, does the subgroup have an ability that if measured and included in selection would both reduce adverse impact for the subgroup and increase validity overall? Can an ability be found which contributes to the measurement of this subgroup's likelihood of success, but is not as important for other subgroups? What would a test battery based solely on predictor and criterion data for just one subgroup look like compared to that of the whole population and for other subgroups? The answers to these questions are unknown but could be determined by extensive research and simulations. The recommendations in this paper include a step in that direction.

4.8 Steps Required to Track Validity with Test Changes

The consideration of adverse impact and validity for subgroups suggests the following sequence of events, including those already completed and those projected:

1. Determine test validity for total population of examinees (completed)
2. Measure possible presence of adverse impact for subgroups separately (completed)
3. Implement test methods to reduce adverse impact that are very unlikely to be detrimental to validity or other subgroups' adverse impact; (not performed, but potential methods are included in this report)

4. Research methods to reduce adverse impact that that may affect validity or be detrimental to other subgroups' adverse impact (begun, but most work remains not performed)
5. Implement only those methods to reduce adverse impact that have been shown to not be detrimental to any group's validity or adverse impact (not performed).

In mitigating adverse impact by modifying the personnel testing process, this sequence of actions is necessary in many cases to adequately understand how the modifications could affect the Air Force and future examinees.

4.9 Adverse Impact at the Item, Test, and Composite Level

Substantial performance differences between subgroups can occur at the item level, subtest level, and composite level. For many decades, developers of Air Force tests have avoided items that were judged by the test developers to have a potential impact on subgroups, including socioeconomic subgroups. Sensitivity reviews of AFOQT items with a verbal component have systematically been conducted for this purpose.

Since personnel decisions are made at the composite level, an examination of composite-level adverse impact would be advisable and likely necessary. To reduce adverse impact for composites, subtest weights could be adjusted if the result were to both reduce adverse impact and maintain validity for all subgroups. A cautionary consideration is that with the many changes that this creates for each subgroup in terms of adverse impact and validity, it is unlikely that a change could be made that was entirely acceptable. This approach with the current subtests is probably better conceptually than practically. If one or more subtests are added to the composite, as would be a consequence of following one of the recommendations of this study, tests added to a composite may both reduce adverse impact and add validity.

4.10 False Flag?

To understand the roots of adverse impact more fully, factors besides race, ethnicity, and sex should be evaluated. Observed adverse impact could be flying under the false flag of race and ethnicity when the underlying factors are instead socioeconomic status, culture, and other factors related to test performance which covary with race and ethnicity. If so, pursuing adverse impact reduction through efforts focused on race and ethnicity may prove of less value. Addressing race when the real issue is socioeconomic status would lead to only indirect solutions, as solutions are

best matched with the true underlying issue. Although, without data representing socioeconomic status and cultural factors, it may not be possible to parse these variables. With data and data analysis of such factors as culture and socioeconomic status, a new categorization of people, cutting across race and ethnicity, may replace race, ethnicity, and sex as categorizations of concern. Although federal law requires race, ethnicity, and sex to be considered, the better improvements may be made only if socioeconomic status and other factors listed are addressed.

4.11 Levels of Effort Required

The recommendations from this study, including those requiring only minor adjustments to the tests and those for adding more tests, would not place a large burden on the test management process. Many possible actions to mitigate adverse impact could require much more attention from managers of personnel testing. New test development policies would be required, and many additional research efforts would be needed to ensure compliance. For example, if a change in the test to reduce adverse impact for Subgroup A resulted in lower validity for that subgroup or any other subgroup, what would be an acceptable tradeoff? Or a reduction of adverse impact for one subgroup which also adds or increases adverse impact for another subgroup? Any strategy or plan to substantially modify tests to reduce adverse impact would need to accommodate all subgroup considerations, thereby significantly increasing the level of effort required.

4.12 Possible Actions, Likely Outcomes, and Comments

(listed in order of ascending advisability)

1. Action:

Reduce or eliminate AFOQT measures.

Outcome:

a. A requirement to develop an alternative selection and classification process
   Likely detrimental to the Air Force in terms of less effective use of human resources
b. Detrimental to those who would have been selected with the test but were not, and would have succeeded if selected
c. Detrimental to those selected who would not otherwise have been selected and subsequently fail in training or on the job
d. Advantage to those who otherwise would not have been selected but were selected and subsequently succeeded
e. A likely reversal of this action if negative consequences become manifest.

Comment:

The course most damaging to the Air Force and examinees with more ability.
Not Recommended.

2. Action:

Perform studies of what a selection and classification process would look like by developing a separate process for each subgroup, then based on the data, make a policy of how to proceed.

Outcome:

a. Long-term extensive research programs along with increased importance of subgroup identity; other outcomes dependent on how the Air Force accommodates the more complex process.
b. Possible complications concerning rules for people of mixed race, although steps recommended here are intended to avoid this problem.

Comment:

This route could be followed with a sample of just a few Air Force job specialties. Not recommended.

3. Action:

Evaluate various non-test practices, such as interviews and resume reviews, following hiring practices found in civilian organizations.

Outcome:

As an adjunct to testing, likely to improve the selection and classification process, with its effectiveness a function of how well it is executed.

Comment:

Labor intensive, adds subjectivity, lacks transparency, and with an unpredictable effect on adverse impact. The process could be evaluated on a small scale, starting with currently problematic training.

4. Action:

In addition to any other course of action concerning the AFOQT, identify constructs that have some potential for validity and at which one or more subgroups would perform well enough to reduce overall adverse impact.

Outcome:

Possibly reducing adverse impact and modestly incrementing validity.

Comment:

High-risk research because the effort might not be successful, but if successful, it could add an innovative method of reducing adverse impact and be significantly beneficial. Recommended if a simulation can demonstrate a benefit and the budget is risk tolerant.

5. Action:

Retain *status quo* for selection and classification testing**.**

Outcome:

The best process for treating the population of examinees as a whole and by far the most efficient for the Air Force due in part to lower rates of training failures.

Comment:

A course that is simple, economical, and acceptably effective, but not enhanced. Recommended if cost is the prime or overriding consideration.

6. Action:

Retain current testing and evaluate constructs identified in recent studies initiated by Strategic Personnel Research Program to determine if they would both improve validity for subgroups and reduce adverse impact.

Outcome:

Unknown but constructs previously recommended for consideration could result in benefits to individuals and the Air Force. It may be difficult to relate to available criteria, but additional tests may have a positive long-term effect on Air Force service performance.

Comment:

The strongest recommendation (see section 3.2.1).

4.13 Finally

Adverse impact is undesirable; a reduction in the effectiveness of the selection and classification process is unacceptable. The suggested goal is to maintain or improve that process while attempting to reduce its adverse impact.

5.0 CONCLUSIONS

The dual benefits of the AFOQT are that the Air Force makes the best use of its human resources and that examinees have an opportunity to demonstrate their qualifications for appropriate training and military occupations. The strength of these dual benefits is a function of the tests' relationships to suitable criteria. A test's value is based on the cornerstone of its validity; therefore, maintaining test effectiveness and the resulting positive impact is a requirement for any effort to reduce adverse impact.

Subgroup differences on the AFOQT subtests are generally like those found on other similar tests, leading to an inference that subgroup differences are not artifacts of the AFOQT.

Retaining the benefits of the AFOQT requires maintaining a process that provides validity to the selection and classification process.

Methods to reduce adverse impact can result in conflicting goals, such as decreasing adverse impact for one subgroup which increases it for another subgroup. New policies would be required to adjudicate the conflicts.

Dealing with the selection and classification process at the subgroup level will require a much larger research and data collection effort than currently exists to determine that no detrimental effects are a result.

Studies from the academic community to investigate adverse mitigation methods have had limited success. The methods we have recommended for evaluation have a wide range of effort required to implement. Recent studies initiated under the Air Force Strategic Personnel Research Program performed to identify tests of attributes and competencies to improve selection and classification testing, would also likely reduce adverse impact. Those tests might support the confluence of the objectives of increased validity and reduced adverse impact, justifying priority consideration.

There will be an Air Force officer selection and classification process. The use of the AFOQT meets the objectives of having validity and demonstrating a lack of bias, but it does not eliminate adverse impact. There are recommended ways of efforts for improving the process, but there is not yet a clear and certain path to simultaneously fully achieving the three objectives of validity, lack of bias, and no adverse impact.

Part 2

Supplementary Information

1.0 EXTENDED INTRODUCTION CONTENT

1.1 Background

1.1.1 AFOQT

This paper describes whether demographic subgroup differences found in the subtest constructs measured by the current operational version of the AFOQT, Form T, align with those found in other measures. It also evaluates methods of ameliorating adverse impact as they relate to Form T. This section presents a review of Form T's content and structure relevant to these evaluations. Form T is composed of 9 subtests. When factor analyses have been conducted for the subtests, the best fitting structures group subtests under these headings: Verbal, Quantitative, Spatial, Aircrew, and Perceptual Speed (see Carretta et al., 2016; Drasgow et al., 2010).

The Verbal subtests measure the ability to comprehend written language. The subtests are Verbal Analogies (VA), Word Knowledge (WK), and Reading Comprehension (RC). Verbal Analogies measures verbal ability and ability to infer implied relationships. The Word Knowledge subtest measures vocabulary acquisition and understanding of language through synonyms. The Reading Comprehension subtest examines the ability to read, and requires the ability to abstract, generalize, and reason constructively.

The Quantitative subtests are Arithmetic Reasoning (AR), which assesses the ability to understand and manipulate relationships to arrive at solutions to word problems; and Math Knowledge (MK), which requires knowledge of mathematical terms and principles.

The Spatial subtest is Block Counting (BC), which assesses spatial ability by requiring examinees to accurately visualize areas they cannot see in a pile of blocks (Drasgow et al., 2010; Carretta et al., 2016). Examinees must count the number of sides a given block has in contact with other blocks. A sample from the AFOQT information pamphlet is included in Figure 1 (United States Air Force, 2015, p. 25). Block 1 in this figure touches three other blocks, making the correct answer for Block 1 B.

KEY					
Block	**A**	**B**	**C**	**D**	**E**
1	2	3	4	5	6
2	5	6	7	8	9
3	1	2	3	4	5
4	3	4	5	6	7
5	2	3	4	5	6

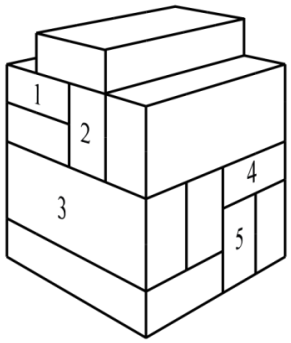

Figure 1. Sample Block Counting items. From *Officer Qualifying Test (AFOQT) Information Pamphlet* (p. 25), by United States Air Force, 2015, Author.

The Perceptual Speed subtest is Table Reading (TR) which measures the ability to extract information from tables (Drasgow et al., 2010; Carretta et al., 2016). A sample of this subtest is included in Figure 2 (United States Air Force, 2015, p. 18). In this sample, the coordinates for Item 1 intersect at a value of 33, making the correct answer D.

		X VALUE						
		-3	-2	-1	0	+1	+2	+3
Y VALUE	+3	25	26	28	30	31	32	33
	+2	26	28	30	32	33	34	35
	+1	27	29	31	33	35	36	37
	0	29	30	32	34	36	37	38
	-1	30	32	33	35	37	38	40
	-2	31	33	34	36	38	39	41
	-3	32	34	35	37	39	40	42

	X	Y	A	B	C	D	E
1.	+1	+2	35	36	30	33	34
2.	0	-3	29	37	39	30	36
3.	-2	+3	26	32	34	28	40
4.	-1	0	33	30	35	36	32
5.	+3	-1	41	27	40	38	39

Figure 2. Sample Table Reading items. From *Officer Qualifying Test AFOQT) Information Pamphlet* (p. 18), by United States Air Force, 2015, Author.

The Aircrew subtests are Instrument Comprehension (IC), Aviation Information (AI), and Physical Science (PS). Instrument Comprehension measures the ability to determine the position of an aircraft according to its pitch, roll, and yaw from illustrations of flight instruments. Figure 3 provides a sample item (United States Air Force, 2015, p. 21). In this item, the instruments to the right indicate that the plane is tilted downwards towards the west and is facing north, making the correct answer A. Aviation Information measures general aeronautical concepts and principles. Physical Science (which replaced the subtest General Science used in previous forms) measures knowledge of basic scientific terms and principles focusing on the physical sciences (Berger et al., 1990; Drasgow et al., 2010; Carretta, et al., 2016).

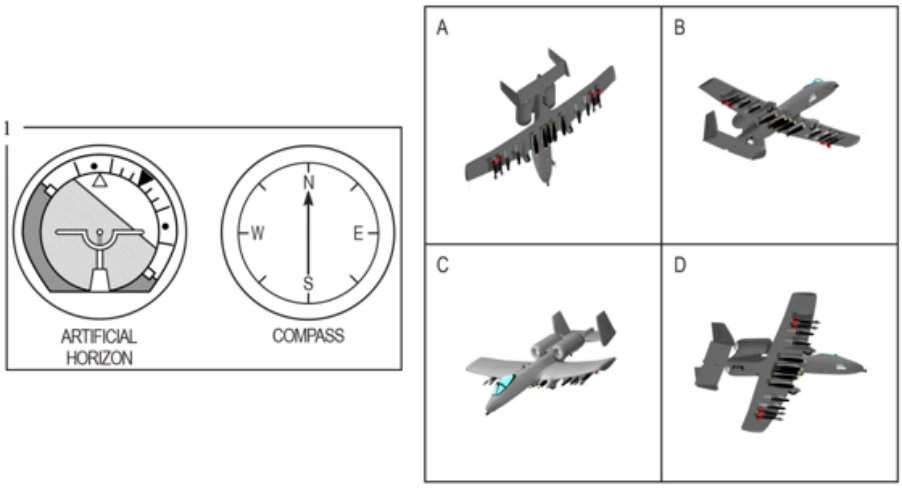

Figure 3. Sample Instrument Comprehension item. From *Officer Qualifying Test (AFOQT) Information Pamphlet* (p. 21), by United States Air Force, 2015, Author.

The subtests are used in 6 operational composites: The Pilot composite, the Combat System Officer (CSO) composite, the Air Battle Manager (ABM) composite, the Academic Aptitude composite, the Verbal composite, and the Quantitative composite. A breakdown of each composite is as follows:

- **Pilot-**Math Knowledge, Table Reading, Instrument Comprehension, and Aviation Information
- **Combat System Officer**-Word Knowledge, Math Knowledge, Table Reading, and Block Counting
- **Air Battle Manager**-Verbal Analogies, Math Knowledge, Table Reading, Instrument Comprehension, Block Counting, and Aviation Information
- **Academic Aptitude**-Verbal Analogies, Arithmetic Reasoning, Word Knowledge, Math Knowledge, and Reading Comprehension
- **Verbal**-Verbal Analogies, Word Knowledge, and Reading Comprehension
- **Quantitative**-Arithmetic Reasoning and Math Knowledge (Drasgow, Nye, Carretta, & Ree, 2010; Carretta, Trent, & Rose, 2016).

Note that Physical Science is not included in any operational composite.

The Verbal, Quantitative, and Academic Aptitude composites are used to qualify applicants to ROTC and OTS, and the Pilot, Combat Systems Officer, and Air Battle Manager (ABM) composites are used to qualify applicants for aircrew training (Carretta, Trent, & Rose, 2016).

Additionally, the AFOQT has shown validity for officer training performance, aircrew training performance criteria, and has shown evidence of validity for several non-aviation officer jobs (see Carretta et al., 2016 referring to Roberts & Skinner, 1996; Carretta, 2008, 2013; Carretta & Ree, 2003; Olea & Ree, 1994; Arth, 1986; Arth & Skinner, 1986; Carretta, 2010; Finegold & Rogers, 1985; Hartke & Short, 1988).

1.1.2 Adverse Impact

Adverse impact was defined by the *Uniform Guidelines on Employee Selection Procedures* (*Guidelines*; Equal Employment Opportunity Commission, 1978) as a substantial difference in selection rates between a majority and minority group by a selection test. Because of laws such as Title VII of the Civil Rights Act (CRA) of 1991 (the CRA was amended in 1991 to explicitly prohibit unjustified adverse impact), adverse impact became a major concern regarding the measurement of underlying characteristics for selection purposes. Adverse impact is unique from other discrimination in that it does not require an intention to discriminate; it consists of an unintentional impact on minorities. Thus, adverse impact *per se* is not unlawful (Equal Employment Opportunity Commission, 1978). Adverse impact is generally only considered discriminatory if the selection procedure is unrelated to performance on the job and business necessity or if another test with lower adverse impact could have reasonably been used instead.

While adverse impact is not unlawful, minimizing it is a desirable goal in its own right. Adverse impact as a phenomenon decreases the proportion of certain groups in an organization. Limiting adverse impact serves to increase diversity and produce an organization that represents the people it serves. However, issues arise when seeking to minimize adverse impact. The primary issue is that the measures which are most valid for predicting performance tend to have the highest adverse impact against minorities: the validity-diversity tradeoff (Ployhart & Holtz, 2008; Pyburn, Ployhart, & Kravitz, 2008). Thus, minimizing adverse impact is difficult to achieve in practice without lowering validity.

1.2 Method and Approach

Articles were searched using Google Scholar, Psych Info, and ERIC. Peer-reviewed research, conference presentations, theses/dissertations, and papers from prominent testing companies were reviewed, focusing on large-scale studies and meta-analyses from 2001 forward. From thousands of articles available on the topic, over 200 articles relating to adverse impact were included in this review, of which 20 presented data only on subgroup differences, and the others discussed methods for reducing adverse impact or important considerations therein. Throughout this report, adverse impact is defined as a subgroup mean difference of at least 0.40 standard deviations, that is, d = 0.40. Used here, "*d"* represents Cohen's *d* effect size, a standardized measure of the difference between two groups. Any discussion of the existence of adverse impact for a measure refers to whether the effect size crosses the threshold of $d = 0.40$. Calculations were always performed such that negative effect sizes indicate the minority/female group scored higher on the construct, and positive effect sizes indicate the majority/male group scored higher on the construct.

2.0 EXTENDED DISCUSSION ON SUBGROUP DIFFERENCES IN THE LITERATURE AND AFOQT

2.1 Approach to Making Construct Comparisons

Comparisons between the subgroup differences for subtests in the AFOQT and subgroup differences found in the literature were made whenever we could find sufficiently similar measures. The extent to which close matches could be made and approaches used for only roughly approximate comparisons are included here. Assessments for verbal and quantitative ability frequently reported data only for composites of multiple specific measures. In cases such as these, comparisons were made with the AFOQT Form T Verbal and Quantitative operational composites instead of a specific subtest. Block Counting had no approximate matches in the literature, so spatial tasks were used in comparisons, with notes on differing content. Instrument Comprehension was also compared to spatial tests, because of the test's similarity to spatial rotation and spatial orientation tests. Instrument Comprehension may rely on knowledge of aircraft instruments, spatial rotation ability, and spatial orientation ability. Note that examinees commonly solve spatial orientation items using spatial rotation strategies (Carpenter & Just, 1986; Carroll, 1993). No similar tests were found regarding Aviation Information. Information on aviation participation is

used as a proxy, as the degree of knowledge regarding aviation principles and instruments is likely related to the degree of participation in aviation. Sufficiently similar tests were found for physical science and perceptual speed so that direct comparisons were made. The personality subtests (i.e., Team Player, Stress Under Pressure, Unassertive, Hyper-Competitive, and Dominance-Leader) were compared to findings in the Big Five personality factors, as well as any Big Five facets or other similar inventories for which findings were available.

2.2 Comparison Considerations

We prioritized large-scale or meta-analytic studies to reduce the influence of sampling error on the resulting subgroup differences. We also used the most recent datasets available and sought samples close in age and education level to the AFOQT's testing population. However, some differences in the samples and populations remain. Some estimates, such as those for the GRE, are based on data from individuals with higher ability, while others, such as the NAEP, offer nationally representative estimates of a subset of the U.S. population (in the case of the NAEP, 12th-grade students). Some samples were conducted on high school students and adolescents, while others were conducted on college students, working adults, or graduate students. Estimates of subgroup differences for those with high ability levels may be larger than the mean values of the population, as differences are exaggerated at the tail ends of normal distributions. Likewise, range restriction in a sample may reduce the estimated effect size. In samples included here, range restriction most frequently occurred due to prior selection. Motivational differences may also influence estimates. For example, several high-stakes tests (i.e., the SAT, GRE, ACT, and MCAT) were used as comparison assessments. While the AFOQT allows two attempts, the SAT allows unlimited attempts, the ACT allows 12, the GRE allows five per year, and the MCAT allows seven. The differences across tests in forgiveness for low test performance may impact the motivation levels of individuals taking the tests. Finally, the types of items and content of items for a given construct varied among comparison tests, and these differences could influence the size of subgroup differences. While these considerations apply primarily to the ability subtests in the AFOQT, there is also a concern particular to the personality measures; a personality trait may not be linearly related to performance, which complicates the interpretation of differences. Other relevant influential factors for effect size estimates are noted when they appear and should be considered

when evaluating the differences between effect sizes found in the literature and those found in the AFOQT.

2.3 Findings

Tables 1 and 2 present a summary of all data found outside the Air Force regarding the achievement constructs measured in the AFOQT. Note that while the following sections include effect sizes found from small-scale studies, these tables include only stable estimates. The SDI-O measures included in the AFOQT are not reported in this table either. For these personality measures, roughly equivalent comparisons for the race and ethnicity subgroups were found only for Stress Under Pressure and Dominance Leader. For female examinees, roughly comparable inventories were found for Stress Under Pressure, Dominance Leader, and Unassertive. These comparison measures, as well as the measures in the AFOQT, mostly showed minimal subgroup differences. A notable exception is Stress Under Pressure. Both the comparison measures in the literature and the AFOQT showed small to moderate differences, favoring either the majority or minority group depending on the measure and group.

Table 1. Comparison of Effect Sizes for AFOQT and Academic Achievement Tests

Subtest	AFOQT				SAT				GRE				ACT		**	MCAT		**
Subgroup*>	**B**	**H**	**A**	**F**	**B**	**H**	**A**	**F**	**B**	**H**	**A**	**F**	**B**	**H**	**A**	**B**	**H**	**A**
Verbal Composite	**0.95**	**0.40**	**0.49**	0.29					**0.94**	**0.56**	0.10	0.36						
Verbal Analogies	**0.86**	0.39	0.34	0.19														
Reading Comprehension	**0.97**	0.33	**0.54**	0.37	**0.95**	**0.77**	-0.01	0.02					**0.86**	**0.55**				
Word Knowledge	**0.77**	0.35	**0.40**	0.24														
Quantitative Composite	**0.85**	0.35	-0.26	**0.46**	**1.04**	**0.77**	**-0.63**	0.25	**0.95**	**0.51**	**-0.41**	**0.59**	**0.90**	**0.51**				
Arithmetic Reasoning	**0.93**	0.35	-0.10	**0.54**														
Math Knowledge	**0.78**	0.30	-0.35	0.39														
Block Counting	**1.03**	0.22	0.15	**0.49**														
Table Reading	**0.82**	0.26	0.10	0.15														
Instrument Comprehension	**1.15**	0.28	**0.40**	**1.08**														
Physical Science	**0.87**	0.28	0.11	**0.66**									**0.97**	**0.61**	0.04	**0.80**	**0.60**	0.29
Aviation Information	**0.88**	0.34	**0.54**	**0.81**														

Note: Effect sizes meeting $d = |0.40|$ are in bold. SAT = Administration in 2016 (College Board, 2016); GRE = All examinees from July 2017 to June 2018; ACT = all examinees from 1997-2005 & 2007; MCAT = Administration in 2009 (Davis et al., 2013).

* B: Black; H: Hispanic; A: Asian; F: Female

** ACT and MCAT administered in 1998 (Camara & Schmidt, 1999).

Table 2. Comparison of Effect Sizes for AFOQT and Other Measures

Subtest	AFOQT				NAEP				Meta-analysis			FAA selection			
Subgroup*>	B	H	A	F	B	H	A	F	B	H	F	B	H	A	F
Verbal Composite	**0.95**	**0.40**	**0.49**	0.29					**0.83**	**0.40**	-0.11				
Verbal Analogies	**0.86**	0.39	0.34	0.19							0.16	**0.83**	**0.74**	0.22	0.20
Reading Comprehension	**0.97**	0.33	**0.54**	0.37	**0.75**	**0.49**	-0.05	-0.25			-0.03				
Word Knowledge	**0.77**	0.35	**0.40**	0.24	**0.82**	**0.63**	0.13	0.00			-0.02				
Quantitative Composite	**0.85**	0.35	-0.26	**0.46**	**0.95**	**0.66**	-0.29	0.09	**0.74**	0.28	0.15				
Arithmetic Reasoning	**0.93**	0.35	-0.10	**0.54**								**1.13**	**0.76**	0.13	**0.72**
Math Knowledge	**0.78**	0.30	-0.35	0.39											
Block Counting	**1.03**	0.22	0.15	**0.49**					**0.66**		**0.57**				
Table Reading	**0.82**	0.26	0.10	0.15							**-0.21, -0.43**	**0.47**	0.28	-0.03	0.06
Instrument Comprehension	**1.15**	0.28	**0.40**	**1.08**					**0.66**		**0.57**				
Physical Science	**0.87**	0.28	0.11	**0.66**	**1.04**	**0.72**	-0.23	0.22							
Aviation Information	**0.88**	0.34	**0.54**	**0.81**											

Note: Effect sizes meeting $d = |0.40|$ are in bold. NAEP = National Assessment of Educational Progress, a national standardized examination given to a probability sample of students in the U.S.—twelfth-grade data from 2015 used here; Meta-analysis = all race/ethnic estimates from Roth et al. (2001) except spatial data (compared to Block Counting and Instrument Comprehension) for Black examinees from Schmitt et al. (1996), Female estimates for verbal constructs from Hyde & Linn, (1988), female data for quantitative constructs from Else-Quest et al. (2010), female data for spatial rotation (compared to Block Counting and Instrument Comprehension) from Maeda & Yoon (2013), female data for perceptual speed (compared to Table Reading) from Hedges & Nowell, (1995); FAA selection = Barrier analysis conducted on the Federal Aviation Administration's selection processes for Air Traffic Control Specialist applicants (Outtz & Hanges, 2013).

* B: Black; H: Hispanic; A: Asian; F: Female

2.3.1 Black-White Differences in Mean Scores

Table 3 provides summary information on the standardized mean differences found between Black and White individuals across the literature reviewed, compared to the values found in the AFOQT. The left side of the table includes the standardized differences for the main constructs and the right side includes any specific subtests within the construct to the left that were available in the literature.

Table 3. Black-White Standardized Differences in the Literature and AFOQT: Aptitude Assessments

Main Construct			Specific Subtests			
			Verbal Analogies	Reading Comprehension	Word Knowledge	Arithmetic Reasoning
Verbal Composite	AFOQT	**0.95**	**0.86**	**0.97**	**0.77**	
	Literature	**0.94**[b], **0.83**[d]	**0.83**[g]	**0.95**[a], **0.86**[c], **0.75**[h]	**0.82**[h]	
Quantitative Composite	AFOQT	**0.85**				**0.93**
	Literature	**1.04**[a], **0.95**[b], **0.90**[c], **0.74**[d], **0.95**[h]				**1.13**[g]
Physical Science	AFOQT	**0.87**				
	Literature	**0.80**[e], **0.97**[c], **1.04**[h]				
Spatial Tests	AFOQT	**1.03** (BC), **1.15** (IC)				
	Literature	**0.66**[f]				
Perceptual Speed	AFOQT	**0.82**				
	Literature	**0.47**[g]				

Note: Positive values indicate that White examinees scored higher, and negative values indicate that Black examinees scored higher. Effect sizes meeting the $d = |0.40|$ threshold are in bold. BC = Block Counting, IC = Instrument Comprehension. For the AFOQT, White $N = 25{,}148$, Black $N = 3{,}308$.

[a]The class of 2016 who took the SAT. Data included for the critical reading and mathematics sections. White $N = 742{,}436$, Black $N = 199{,}306$ (College Board, 2016).

[b]U.S. citizens who took the GRE from July 2017-June 2018 using most recent test scores. Data included for the verbal reasoning and quantitative reasoning sections. Black $N = 26{,}665$, White $N = 182{,}623$ (Educational Testing Service, 2018).

[c]All White and Black examinees for the ACT over 1997-2005, and 2007 (Sackett & Shen, 2010).

[d]Meta-analytic estimate from applicant industrial samples (Roth et al., 2001).

[e]Examinees who took the MCAT in 2009 using their most recent test scores. White $N = 33{,}807$, Black $N = 6{,}183$ (Davis et al., 2013).

[f]Meta-analytic estimate of spatial ability (Schmitt et al., 1996).

[g]Sample of applicants taking tests administered by the Federal Aviation Administration for Air Traffic Control Specialist selection from 2006-2011. White $N = 10{,}035$, Black $N = 3{,}197$ (Outtz & Hanges, 2013).

[h]NAEP National data for 12th-grade students from the year 2015 for physical science, the mathematics composite, the reading comprehension composite, and the meaning vocabulary scale.

A. Verbal and Quantitative Reasoning. For verbal and quantitative abilities, comparisons were made against other high-stakes standardized tests (i.e., the SAT, GRE, and ACT) as well as with an extensive meta-analysis (i.e., Roth et al., 2001) which included individuals 14 years or older, and data from the National Assessment of Educational Progress (NAEP) for the 12th grade from 2015. The SAT, GRE, and ACT all include essay sections, but data for these sections were not included anywhere because the AFOQT has no writing section. Notably, the NAEP makes efforts to produce a nationally representative sample (e.g., they use a probability sampling method), and offers accommodations for individuals who are English Language Learners (ELLs) or have a disability. The Roth et al. meta-analysis included data from educational tests, military samples, and industrial samples. Unless noted otherwise, when referring to this meta-analysis, we include only effect sizes from industrial samples. For

industrial samples, Roth et al. (2001) calculated effect sizes separately for those who were incumbents or applicants, from which we include applicant samples only.

Effect sizes for verbal skills found in the literature were very similar to those in the AFOQT; differences between effect sizes for the AFOQT subtests and those of other measures generally did not exceed 0.20 standard deviations. Effect sizes were also all in the same direction, with White examinees scoring higher than Black examinees. The GRE (Educational Testing Service, 2018) and a meta-analysis (Roth et al., 2001) were compared to the Verbal composite in the AFOQT, both of which had almost equivalent effect sizes to the AFOQT. For this subgroup, specific comparisons were found for Word Knowledge (in the NAEP 2015), Reading Comprehension (in the NAEP 2015, SAT; College Board, 2016, and ACT; Sackett & Shen, 2010), and Verbal Analogies (in a barrier analysis for the FAA on air traffic controller applicants; Outtz & Hanges, 2013). The differences found in the literature for these subtests were also mostly equivalent to those found in the AFOQT. The largest difference between the AFOQT and other measures for these subtests was found for the NAEP 2015 reading comprehension measure, which was 0.23 standard deviations smaller than the difference in the AFOQT.

Quantitative measures in the literature showed the same similarity in magnitude and direction as those for the verbal measures. Comparisons for the AFOQT's Quantitative composite were made against the SAT (College Board, 2016), ACT, (Sackett & Shen, 2010), GRE (Educational Testing Service, 2018), NAEP 2015, and meta-analysis (Roth et al., 2001).

Only the barrier analysis conducted for the FAA was found to have a test sufficiently similar for direct comparison to one of the AFOQT's quantitative subtests, Arithmetic Reasoning. The FAA's applied math test consisted of face valid items for speed, time, and distance (Outtz & Hanges, 2013), which are similar to the word problems in the AFOQT's Arithmetic Reasoning. These effect sizes showed a similar White advantage, with the effect size found in the literature producing the larger difference by 0.20 standard deviations.

Notably, Roth et al. (2001) found that overall, Black-White effect sizes were larger for educational samples (i.e., those for the SAT, GRE, and ACT) than for industrial samples, largely due to differences on the GRE. In our more recent effect sizes for educational tests, the GRE has not maintained a disproportionate effect size. The difference was larger for the military samples as well. The authors noted these studies were more likely to represent the entire U.S. population. Roth et al also found larger differences for less complex jobs, suggesting that hiring rates should be more equal for jobs of higher complexity.

B. Physical Science. The Medical College Admission Test (MCAT) included a physical science subtest until April 2015, when the test content was updated to include chemistry and be more directly applicable to medical physics (Beck, 2015). MCAT Data from 2009 is used here. The MCAT is comparable to the AFOQT in that it is a high-stakes selection test. The test likely differs from the AFOQT in the subset of the population taking the exam, as the MCAT is probably taken primarily by very high-ability individuals. An analysis of the MCAT data found a subgroup difference approximately equal in direction and magnitude to that in the AFOQT (Davis et al., 2013). The NAEP 2015 included a physical science test, for which results represent high school students across the United States. The ACT Science test is also comparative to the AFOQT's Physical Science test, although broader. An average across years (1997 through 2005 and 2007) for the ACT Science test (Sackett & Shen, 2010), and the NAEP physical science measure also found differences approximately equal in size and magnitude to that in the AFOQT.

C. Spatial Abilities and Perceptual Speed. An outdated meta-analysis on spatial abilities was the only comparison found for the AFOQT's Block Counting and Instrument Comprehension subtests. While this estimate also showed a White advantage, its magnitude was smaller than the effect sizes for both Block Counting and Instrument Comprehension.

In terms of perceptual speed, the barrier analysis conducted for the FAA included a perceptual speed test; a Scan test, where examinees surveyed moving blocks of data with the aim of identifying numbers outside a predetermined range (Outtz & Hanges, 2013). The subgroup

difference found for this test was $d = 0.47$. While in the same direction, this difference is smaller than that found in the AFOQT for perceptual speed (i.e., Table Reading).

D. Personality Measures. Table 4 provides a summary of the effect size estimates for the Big Five Factors as reported in the literature, in addition to any specific personality facets approximate to the measures used in the AFOQT.

Table 4. Black-White Standardized Differences in the Literature and AFOQT: Personality Measures

Literature Measure	*d*	Black *N*	White *N*	AFOQT Test	*d*	Black *N*	White *N*
Low Anxiety	0.23	359	1,521	Stress Under Pressure	-0.05	3,306	25,143
Even-Tempered	-0.06	806	2,685				
Dominance	0.03	5,214	34,338	Dominance-Leader	-0.02	3,306	25,143
Openness	0.10	3,208	21,749				
Conscientiousness	-0.07	19,195	161,283				
Extraversion	0.16	19,330	90,772				
Agreeableness	0.03	3,297	21,590				
Emotional Stability	0.09	49,719	102,716				

Note: Positive values indicate that White examinees scored higher, and negative values indicate that Black examinees scored higher. Effect sizes meeting the $d = |0.40|$ threshold are in bold. Effect size estimates not drawn from the AFOQT were all drawn from a meta-analytic estimate of adults in the U.S. (Foldes et al., 2008).

Racial and ethnic differences in the Big Five Factors of personality were calculated in a meta-analysis on adults in the U.S. (Foldes, Duehr, & Ones, 2008). Black-White differences were between $d = 0.00$ and $d = \pm 0.10$ for Openness to Experience, Conscientiousness, Agreeableness, and Emotional Stability. Extraversion showed the largest group difference at $d = 0.16$. Despite the minimal reported differences for Emotional Stability, the measure of its facet Low-Anxiety was reported at $d = 0.23$. Low-Anxiety in the Foldes et al. meta-analysis is comparable to Stress Under Pressure in the AFOQT. While low-Anxiety showed a slight difference in favor of White individuals, Stress Under Pressure showed no palpable difference between Black and White individuals. The measure Even Tempered in the meta-analysis was much more like Stress Under Pressure, also showing no difference between Black and White individuals. Note that Stress Under Pressure is negatively coded compared to Even Tempered and Low Anxiety. Dominance, the facet included in the meta-analysis closest to the AFOQT's measure of Dominance-Leader, had a negligible mean difference, as did Dominance-Leader.

2.3.2 Hispanic/Non-Hispanic Differences in Mean Scores

Table 5 provides summary information on the standardized mean differences found between Hispanic and White or Hispanic and non-Hispanic individuals across the literature reviewed, compared to the values found in the AFOQT. The left side of the table includes the standardized differences for the main constructs and the right side includes any specific subtests within the construct to the left that were available in the literature.

Table 5. Hispanic/non-Hispanic Standardized Differences in the Literature and AFOQT: Aptitude Assessments

Main Construct			Specific Subtests			
			Verbal Analogies	Reading Comprehension	Word Knowledge	Arithmetic Reasoning
Verbal Composite	AFOQT	**0.40**	0.39	0.33	0.35	
	Literature	**0.56**[b], **0.40**[d]	**0.74**[f]	**0.77**[a], **0.55**[c], **0.49**[g]	**0.63**[g]	
Quantitative Composite	AFOQT	0.35				0.35
	Literature	**0.77**[a], **0.51**[bc], 0.28[d], **0.66**[g]				**0.76**[f]
Physical Science	AFOQT	0.28				
	Literature	**0.60**[e], **0.61**[c], **0.72**[g]				
Spatial Tests	AFOQT	0.22 (BC), 0.28 (IC)				
	Literature	-				
Perceptual Speed	AFOQT	0.26				
	Literature	0.28[f]				

Note: Positive values indicate that non-Hispanic/White examinees scored higher, and negative values indicate that Hispanic examinees scored higher. Effect sizes meeting the $d = |0.40|$ threshold is in bold. For the AFOQT, non-Hispanic $N = 33{,}580$, Hispanic $N = 5731$.

[a]The class of 2016 who took the SAT. Data included for the critical reading and mathematics sections. White $N = 742{,}436$, Hispanic $N = 355{,}829$ (College Board, 2016).

[b]U.S. citizens who took the GRE from July 2017-June 2018 using their most recent test scores. Data included for the verbal reasoning and quantitative reasoning sections. Hispanic $N = 30{,}539$, White $N = 182{,}623$ (Educational Testing Service, 2018).

[c]All White and Hispanic examinees for the ACT over 1997-2005, and 2007(Sackett & Shen, 2010).

[d]Meta-analytic estimate from industrial samples. $N = 6{,}133$ (Roth et al., 2001).

[e]Examinees who took the MCAT in 2009 using their most recent test scores. White $N = 33{,}807$, Hispanic $N = 5{,}810$ (Davis et al., 2013).

[f]Sample of applicants taking tests administered by the Federal Aviation Administration for Air Traffic Control Specialist selection from 2006-2011. White $N = 10{,}035$, Hispanic $N = 940$ (Outtz & Hanges, 2013).

[g]NAEP National data for 12th-grade students from the year 2015 for physical science, the mathematics composite, the reading comprehension composite, and the meaning vocabulary test.

A. **Verbal and Quantitative Reasoning.** For non-Hispanic/Hispanic differences, comparisons for verbal abilities were made against the same high-stakes tests, meta-analysis, and NAEP data used for Black-White difference comparisons. Note that the analyses conducted for the achievement tests, meta-analysis, barrier analysis, and the NAEP calculated subgroup differences for the Hispanic sample against the White sample, while the AFOQT allows individuals who identify themselves as White to also identify themselves as Hispanic. Effect sizes in the literature were all in the same direction as the AFOQT (i.e., Hispanic individuals had lower scores on average) but tended to be slightly to moderately larger than those found in the AFOQT. The largest gap between the AFOQT and other measures for verbal abilities was found for the SAT reading comprehension test, which had a difference 0.44 standard deviations larger than that in the AFOQT. The Roth et al. (2001) meta-analysis found a verbal reasoning difference of exactly the same magnitude and size as the AFOQT's verbal composite, but they noted that not many industrial samples were available for this analysis (k = 7), and thus there were inadequate data to separate applicant and incumbent samples. Incumbent samples in their meta-analysis produced smaller subgroup differences than applicant samples (Roth et al., 2001).

Quantitative Reasoning comparisons here were also made against the same measures as used for Black-White comparisons. As with the verbal measures, quantitative measures found in the literature also showed Hispanic individuals scoring lower on average but tended to produce larger subgroup differences than those in the AFOQT. The largest of these disparities was between the SAT mathematics section and the Quantitative composite in the AFOQT; the subgroup difference for the SAT was 0.42 standard deviations larger than that of the AFOQT. The Roth et al. (2001) meta-analysis was the only estimate in the literature that found a difference smaller than that in the AFOQT, as its quantitative reasoning estimate was d = 0.28 compared to the AFOQT's Quantitative composite difference of d = 0.35. As with the verbal estimate, Roth et al. did not have enough samples to separate applicants and incumbents (k = 7). Once again, the discrepancies observed here may be partially accounted for by the difference in group demarcation.

B. Physical Science. As with the verbal and quantitative subtests, the comparable science measures here are the same as those used in the Black-White section for Physical Science. Differences found in the literature, while showing the same Hispanic disadvantage as the effect size in the AFOQT, were around 0.30 to 0.40 standard deviations larger than that in the AFOQT. The largest disparity was found in comparison to the NAEP estimate, which was 0.44 standard deviations larger.

C. Spatial Abilities and Perceptual Speed. No appropriate spatial tests were found to use as comparisons to the AFOQT subtests. The Scan subtest difference calculated in the ATCS barrier analysis (Outtz & Hanges, 2013) was approximately equal to the AFOQT perceptual speed measure Table Reading in both size and magnitude.

D. Personality Measures. Table 6 provides a summary of the effect size estimates for the Big Five Dimensions as reported in the literature, in addition to any specific personality facets approximate to the measures used in the AFOQT.

Table 6. Hispanic/non-Hispanic Standardized Differences in the Literature and AFOQT: Personality Measures

Literature Measure	*d*	Hispanic *N*	White *N*	AFOQT Test	*d*	Hispanic *N*	Non-Hispanic *N*
Low Anxiety	-0.25	206	806	Stress Under Pressure	0.04	5,729	33,573
Even-Tempered	-0.09	299	2,060				
Dominance	0.04	4,376	30,615	Dominance-Leader	-0.04	5,729	33,573
Openness	0.02	3,082	21,911				
Conscientiousness	-0.08	81,564	151,207				
Extraversion	0.02	20,449	74,071				
Agreeableness	0.05	3,052	21,588				
Emotional Stability	-0.03	28,327	95,754				

Note: Effect sizes meeting the $d = |0.40|$ threshold are in bold. Positive effect sizes indicate that non-Hispanic/White individuals scored higher, and negative effect sizes indicate that Hispanic individuals scored higher. Effect size estimates not drawn from the AFOQT were all drawn from a meta-analytic estimate of adults in the U.S. (Foldes et al., 2008).

The Foldes et al. (2008) meta-analysis found that for the Big Five Dimensions, there was no difference between Hispanic and White individuals (all effect sizes were under ±0.10 in magnitude). In terms of the facets most similar to the personality traits measured in the AFOQT, Low Anxiety had an effect size favoring Hispanic individuals, compared to no

palpable difference found for Stress Under Pressure in the AFOQT. Even Tempered produced an estimate more similar to the AFOQT estimate. Note that Stress Under Pressure is negatively coded compared to Even Tempered and Low Anxiety. Neither Dominance in the meta-analysis nor Dominance-Leader in the AFOQT showed any palpable difference between the groups.

2.3.3 Asian-White differences in Mean Scores

Table 7 provides summary information on the standardized mean differences found between Asian and White individuals across the literature reviewed, compared to the values found in the AFOQT. The left side of the table includes the standardized differences for the main constructs and the right side includes any specific subtests within the construct to the left that were available in the literature.

Table 7. Asian-White Standardized Differences in the Literature and AFOQT: Aptitude Assessments

Main Construct			Specific Subtests			
			Verbal Analogies	Reading Comprehension	Word Knowledge	Arithmetic Reasoning
Verbal Composite	AFOQT	**0.49**	0.34	**0.54**	**0.40**	
	Literature	0.10[b]	0.22[j]	-0.01[a], -0.05[k]	0.13[k]	
Quantitative Composite	AFOQT	-0.26				-0.10
	Literature	**-0.63**[a], **-0.41**[b], -0.29[k]				0.13[j]
Physical Science	AFOQT	0.11				
	Literature	0.29[c], 0.04[d], -0.23[k]				
Spatial Tests	AFOQT	0.15 (BC), **0.40** (IC)				
	Literature	**0.79**[e], 0.34[f], 0.14[g]				
Perceptual Speed	AFOQT	0.10				
	Literature	0.12[h], **0.82**[i], -0.03[j]				

Note: Positive values indicate that White examinees scored higher, and negative values indicate that Asian examinees scored higher. Effect sizes meeting the $d = |0.40|$ threshold are in bold. BC = Block Counting, IC = Instrument Comprehension. For the AFOQT, White $N = 25{,}148$, Asian $N = 1{,}894$.

[a]The class of 2016 who took the SAT. Data included for the critical reading and mathematics sections. White $N = 742{,}436$, Asian $N = 196{,}735$ (College Board, 2016).

[b]U.S. citizens who took the GRE from July 2017-June 2018 using their most recent test scores. Data included for the verbal reasoning and quantitative reasoning sections. Asian $N = 22{,}567$, White $N = 182{,}623$ (Educational Testing Service, 2018).

[c]Individuals who took the MCAT Physical Sciences test in 1998 (Camara & Schmidt, 1999).

[d]Individuals who took the ACT Science test in 1998 (Camara & Schmidt, 1999).

[e]Sample of young adults in the U.S. (mean age = 20.1) and China (mean age = 19.4) who took the Card Rotations test. Chinese $N = 40$, American $N = 66$ (Geary et al., 1996).

[f]Sample of young adults in the U.S. (mean age = 20.1) and China (mean age = 19.4) who took the Cube Rotations test. Chinese $N = 40$, American $N = 66$ (Geary et al., 1996).

[g]Sample of young adults in the U.S. (mean age = 20.1) and China (mean age = 19.4) who took the Mental Rotations test. Chinese $N = 40$, American $N = 66$ (Geary et al., 1996).
[h]Sample of young adults in the U.S. (mean age = 20.1) and China (mean age = 19.4) who took the Number Comparisons test. Chinese $N = 40$, American $N = 66$ (Geary et al., 1996).
[i]Sample of young adults in the U.S. (mean age = 20.1) and China (mean age = 19.4) who took the Identical Pictures test. Chinese $N = 40$, American $N = 66$ (Geary et al., 1996).
[j]Sample of applicants taking tests administered by the Federal Aviation Administration for Air Traffic Control Specialist selection from 2006-2011. White $N = 10{,}035$, Asian $N = 517$ (Outtz & Hanges, 2013).
[k]NAEP National data for 12th-grade students from the year 2015 for physical science, the mathematics composite, the reading comprehension composite, and the meaning vocabulary test.

A. Verbal and Quantitative Reasoning. For the Verbal composite, a comparison was made against the GRE (Educational Testing Service, 2018). Reading Comprehension in the AFOQT was compared to the critical reading test in the SAT (College Board, 2016) and NAEP reading comprehension measure. Word Knowledge was compared to the vocabulary measure in the NAEP. The FAA selection process verbal analogies test (Outtz & Hanges, 2013) was used as a comparison for Verbal Analogies in the AFOQT. The effect sizes for verbal measures in the literature were all smaller than those in the AFOQT. All the verbal subtests in the AFOQT showed a White advantage, while several in the literature showed no palpable difference or only a slight White advantage. The largest disparity was between the reading comprehension effect sizes in the AFOQT and NAEP 2015, which were 0.59 standard deviations apart, with the NAEP measure showing no palpable difference between groups and the AFOQT showing a medium White advantage. The smallest disparity was between the verbal analogies subtests in the AFOQT and the FAA selection process for Air Traffic Controller Specialists, which were 0.12 standard deviations apart, with the AFOQT showing a larger difference in favor of White examinees.

The AFOQT Quantitative composite was compared to the SAT mathematics section (College Board, 2016), GRE quantitative section (Educational Testing Service, 2018), and the NAEP mathematics composite. All of these indices and the AFOQT showed an Asian advantage. The SAT had an effect size larger than the AFOQT, the GRE had an effect size slightly larger than the AFOQT, and the NAEP had an effect size equivalent to that of the AFOQT. In the SAT, Asian examinees' scores in critical reading and mathematics have improved disproportionately to many other groups over a twenty-year period (Kobrin et al., 2006). Arithmetic Reasoning in the AFOQT was compared to the applied mathematics test in the FAA Air Traffic Controller Specialist selection process (Outtz & Hanges, 2013). The magnitudes of these effect sizes were

minimal and similar, but while the applied math test showed White examinees scoring higher, Arithmetic Reasoning in the AFOQT showed Asian examinees scoring higher.

B. Physical Science. Little comparative data have been published in this area. The most appropriate publication found reported data for subgroup differences between Asian and White examinees on the MCAT Physical Sciences subtest in 1998 and the ACT Science Reasoning subtest in 1998 (Camara & Schmidt, 1999). Compared to the AFOQT's effect size of $d = 0.11$, the MCAT Physical Sciences subtest produced a subgroup difference of $d = 0.29$, and the ACT sciences subtest produced a difference of $d = 0.04$. The NAEP from 2015 found a difference of $d = -0.23$ in favor of Asian individuals on physical science. Note that the populations taking these four exams differ.

C. Spatial Abilities and Perceptual Speed. Very little data were available for subgroup differences between White and Asian individuals. One exception is a study conducted by Geary, Salthouse, Chen, and Fan (1996), which administered spatial and perceptual speed tests to American and Chinese adults. The sample size of this study was relatively small, with 110 participants from the U.S. and 80 participants from China. Roughly half of each sample was composed of older adults (note that only data on young adults are included in Table 7). Spatial ability was measured in Geary et al. (1996) with the Card Rotations Test, the Cube Comparisons Test, and the Mental Rotation Test. The Card Rotations Test requires examinees to rotate figures in two dimensions and the Mental Rotation Test requires rotation in three dimensions. Cube Comparisons included both two-dimensional and three-dimensional rotation requirements. These tests were speeded. For the younger adults, the differences between American and Chinese individuals were $d = 0.79$ for the Card Rotations test, 0.34 for the Cube Rotations test, and 0.14 for the Mental Rotation test. Block Counting in the AFOQT produced a mean standardized difference of 0.15, which was in the same direction but smaller than all the spatial ability measures in the Geary et al. study except Mental Rotations. Perceptual speed was measured in the Geary et al. study with the Number Comparison test ($d = 0.12$), in which pairs of digit strings are presented to the examinee, who must determine if the strings are identical. Perceptual speed was also measured by Geary et al. with the Identical Pictures Test ($d = 0.82$), in which examinees must choose which of five responses contains an image identical to a stimulus image. Both tests were also speeded. While these tests both showed an

American advantage, the mean differences for these tests differed greatly (by 0.70 standard deviations). Of the two, the Number Comparison test had the effect size more similar to the difference in the AFOQT's Table Reading subtest (d = 0.10). The similarity is reasonable conceptually, as both the Number Comparison and Table Reading subtest deal with perceptual speed in digits, while the Identical Pictures test does not. Finally, the Scan subtest in the ATCS barrier assessment produced an observed difference of d = -0.03, a negligible effect size. The ATCS estimate is the best here, as it compared individuals on self-reported race instead of nationality, constituted a larger sample over a longer period, and is more recent. The other estimates are rough comparisons at best.

Instrument Comprehension was closest to the three-dimensional spatial rotation test (the Mental Rotations test) in the Geary et al. (1996) study, which produced a difference of d = 0.14. The Instrument Comprehension test had an effect size in the same direction but larger than this (d = 0.40), potentially because it includes variance that the Mental Rotations test did not (e.g., knowledge of aviation instruments).

D. Personality Measures. Table 8 provides a summary of the effect size estimates for the Big Five Dimensions as reported in the literature, in addition to any specific personality facets approximate to the measures used in the AFOQT.

Table 8. Asian-White Standardized Differences in the Literature and AFOQT: Personality Measures

Literature Measure	*d*	Asian *N*	White *N*	AFOQT Test	*d*	Asian *N*	White *N*
Low Anxiety	-0.27	80	728	Stress Under Pressure	**-0.40**	1,894	25,143
Even-Tempered	0.38	882	10,072				
Dominance	0.19	961	15,142	Dominance-Leader	0.18	1,894	25,143
Openness	-0.11	132	1,464				
Conscientiousness	-0.11	3,454	104,257				
Extraversion	0.14	3,013	53,254				
Agreeableness	**-0.63**	93	1,216				
Emotional Stability	0.12	3,398	82,187				

Note: Positive values indicate that White individuals scored higher, and negative values indicate that Asian individuals scored higher. Effect sizes meeting the $d = |0.40|$ threshold are in bold. Effect size estimates not drawn from the AFOQT were all drawn from a meta-analytic estimate of adults in the U.S. (Foldes et al., 2008).

As with the sections for the Black and Hispanic subgroups, the meta-analysis by Foldes et al. (2008) was used as a comparison for personality traits measured in the AFOQT. For the Big Five Dimensions, Emotional Stability (d = 0.12) and Extraversion (d = 0.14) had minimal differences with White individuals scoring higher, Openness to Experience (d = -0.11) and Conscientiousness (d = -0.11) both had minimal difference with Asian individuals scoring higher, and Agreeableness (d = -0.63) had a medium to large difference with Asian individuals scoring higher. In terms of the facets that map most closely to the measures in the AFOQT, Low Anxiety was found by Foldes et al. to have an Asian-White difference of d = -0.27, which is both in the opposite direction and of smaller magnitude than the effect size of d = -0.40 for the AFOQT's Stress Under Pressure (note that Stress Under Pressure is negatively coded, meaning it showed a White advantage while Low Anxiety showed an Asian advantage). Another similar measure, Even Tempered, was found in the Foldes et al. meta-analysis to be in the same direction as Stress Under Pressure and closer in magnitude (d = 0.38). Dominance from the Foldes et al. meta-analysis had the same mean subgroup difference as Dominance-Leader in the AFOQT in both magnitude and direction.

2.3.4 Female-Male Differences in Mean Scores

Table 9 provides summary information on the standardized mean differences found between female and male individuals across the literature reviewed, compared to the values found in the AFOQT. The left side of the table includes the standardized differences for the main constructs and the right side includes any specific subtests within the construct to the left that were available in the literature.

Table 9. Female-Male Standardized Differences in the Literature and AFOQT: Aptitude Assessments

Main Construct			Specific Subtests			
			Verbal Analogies	Reading Comprehension	Word Knowledge	Arithmetic Reasoning
Verbal Composite	AFOQT	0.29	0.19	0.37	0.24	
	Literature	0.36[b], -0.11[c]	0.16[c], 0.20[i]	0.02[a], -0.03[c], -0.25[f]	-0.02[c], 0.00[f]	
Quantitative Composite	AFOQT	**0.46**				**0.54**
	Literature	0.25[a], **0.59**[b], 0.06[d], 0.15[e], 0.09[f]				**0.72**[i]
Physical Science	AFOQT	**0.66**				
	Literature	0.22[f]				
Spatial Tests	AFOQT	**0.49** (BC), **1.08** (IC)				
	Literature	**0.57**[g]				
Perceptual Speed	AFOQT	0.15				
	Literature	-0.21 or **-0.43**[h], 0.06[i]				

Note: Positive values indicate that male examinees scored higher, and negative values indicate that female examinees scored higher. Effect sizes meeting the $d = |0.40|$ threshold are in bold. BC = Block Counting, IC = Instrument Comprehension. For the AFOQT, male $N = 29{,}536$, female $N = 10{,}550$.

[a]The class of 2016 who took the SAT. Data included for the critical reading and mathematics sections. Male $N = 762{,}247$, female $N = 875{,}342$ (College Board, 2016).

[b]U.S. citizens who took the GRE from July 2017-June 2018 using their most recent test scores. Data included for the verbal reasoning and quantitative reasoning sections. Male $N = 113{,}925$, female $N = 199{,}698$ (Educational Testing Service, 2018).

[c]Meta-analytic review of verbal abilities of individuals from under 4 years old to 26 years and older in the U.S. and Canada for overall verbal ability, vocabulary, and reading comprehension. Total $N = 1{,}418{,}899$ (Hyde & Linn, 1988).

[d]Effect sizes for math performance of grade 11 students calculated from records required under the No Child Left Behind legislation from 10 states, where the mean performance for these states was found to match the national mean. Total $N = 446{,}381$ (Hyde et al., 2008).

[e]Meta-analytic summary of mathematic performance for students aged ~14-16 years across 69 countries (Else-Quest et al., 2010).

[f]NAEP National data for 12th-grade students in the year 2015 for physical science, the mathematics composite, the reading comprehension composite, and the meaning vocabulary test.

[g]Meta-analytic review of gender differences in the Purdue Spatial Visualization Tests: Visualization of Rotations (Maeda & Yoon, 2013).

[h]Meta-analytic review of 4 large scale probability samples of adolescents and young adults the U.S. Total $N = 127{,}268$ (Hedges & Nowell, 1995).

[i]Applicants taking tests administered by the Federal Aviation Administration for Air Traffic Control Specialist selection from 2006-2011. Male $N = 11{,}813$, female $N = 3{,}635$ (Outtz & Hanges, 2013).

A. Verbal and Quantitative Reasoning. For the Verbal composite of the AFOQT, comparisons were made against the GRE's verbal reasoning section (Educational Testing Service, 2018) and an outdated meta-analysis (i.e., Hyde & Linn, 1988). This meta-analysis assimilated studies from the U.S. and Canada. It included participants from under five years of age to over 26. Despite the wide age range, it was still deemed appropriate for use here because the researchers generally did not find appreciable differences in effect sizes across ages (Hyde &

Linn, 1988). Like the AFOQT, the GRE showed a small to medium male advantage. The meta-analysis had a slight female advantage.

For specific verbal subtests, Verbal Analogies was compared to an estimate from the meta-analysis (Hyde & Linn, 1988) and a verbal analogies measure in the FAA barrier analysis (Outtz & Hanges, 2013). Reading Comprehension was compared to the SAT's critical reading section (College Board, 2016), an estimate from the meta-analysis (Hyde & Linn, 1988), and the NAEP's reading comprehension composite. Word Knowledge in the AFOQT was compared to an estimate from the meta-analysis (Hyde & Linn, 1988) and the NAEP vocabulary test. Verbal Analogies as measured in the AFOQT and the two measures found in the literature all showed a small male advantage. While the AFOQT showed a small to medium male advantage in Reading Comprehension, effect sizes found in the literature showed no difference or a small female advantage (in the case of the NAEP 2015). The difference of d = 0.02 on reading comprehension in the SAT is notable. On the SAT, women once held an advantage of 5 points in the verbal section, but the difference decreased over time and was lost by 1972 (Kobrin et al., 2006; Lowen et al., 1988). Word Knowledge had a similar pattern. The AFOQT showed a small male advantage and neither measures in the literature found a difference between men and women. Note that in the Hyde and Linn (1988) meta-analysis, women showed the largest advantage in speech production (d = -0.33), a measure not included in the AFOQT.

The Quantitative composite in the AFOQT was compared to the SAT (College Board, 2016), the GRE (Educational Testing Service, 2018), data collected for 11th-grade students from No Child Left Behind legislation (Hyde et al., 2008), a cross-national meta-analysis of 69 nations for students approximately 14 to 16 years old (Else-Quest et al., 2010), and the 12th-grade NAEP data for their mathematics composite. The AFOQT showed a medium male advantage. A larger male advantage was shown in the GRE. The other tests showed no palpable difference to a small male advantage.

As with the racial and ethnic groups, comparisons with specific quantitative subtests were only possible for Arithmetic Reasoning. The FAA's barrier analysis included an applied mathematics subtest that produced a larger male advantage than that in Arithmetic Reasoning. No other evidence for the other quantitative subtest was found, and the comparisons shown here may not be generalizable to it. The Else-Quest et al. (2010) meta-analysis found that the effect sizes for the U.S. varied depending on the quantitative subject area (from d = -0.01 to 0.15).

B. Physical Science. The most recent large-scale study found measuring the difference in physical science performance between men and women was the NAEP 2015, which found a male advantage of $d = 0.22$ for physical science in 2015. This estimate is substantially smaller than the effect size of $d = 0.66$ found in the AFOQT. A possible explanation could be that the general physical science abilities of individuals taking the AFOQT are higher than average, as past studies have found larger differences at the high end of the distribution (Hedges & Nowell, 1995). The higher education level of those taking the AFOQT suggests that there may be such a difference, although we were unable to confirm, given the noncomparability of the measures. Ganley, Vasilyeva, and Dulaney (2014) conducted a study on 73,245 students 13 to 15 years old and found that a $d = 0.14$ difference in physical science achievement was moderated by differences in spatial rotation ability (i.e., items which required greater spatial ability had larger gender differences). The researchers also noted that in a smaller sample from a school that performed higher than the state average, the male-female difference was larger ($d = 0.48$).

C. Spatial Abilities and Perceptual Speed. A meta-analysis on spatial rotation ability found a difference of $d = 0.57$ between men and women on the skill (Maeda & Yoon, 2013). Specifically, the meta-analysis investigated the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R), where the stem includes a figure, and the examinee must identify which response option contains the stem figure rotated in the direction indicated by the instructions. The effect size found by Maeda and Yoon is in the same direction but smaller in magnitude compared to Instrument Comprehension. The differing content of these tests should not be overlooked when comparing their effect sizes. Instrument Comprehension's format is similar to the PSVT:R; examinees view instruments in the stem and then choose the plane rotated in the position indicated by the instruments. The main difference between these

two tests is the stem. In Instrument Comprehension, instruments shown from the plane's perspective must be understood, instead of a block. The block used in the PSVT:R does not require prior knowledge beyond test instructions or involve spatial orientation, but the instruments in Instrument Comprehension do. The requirement of prior knowledge is likely the test content difference accountable for the effect size difference, as Barron and Rose (2013) and Bosco, Longoni, and Vecchi (2004) found a relatively small gender difference in spatial orientation ability. Additionally, research has found that spatial orientation items are frequently solved using mental rotation (Carpenter & Just, 1986; Carroll, 1993), so content differences in orientation versus rotation may not be as profound as they appear.

The Block Counting subtest is also roughly comparable to the PSVT:R in that it requires visualization of a three-dimensional object. At d = 0.49, it has an effect size in the same direction and of similar magnitude to that found by Maeda and Yoon (2013). Note that spatial rotation has been found to have a greater male-female difference than other spatial sub-abilities (Kaufman, 2007; Nordvik & Amponsah, 1998; Voyer, Voyer, & Bryden, 1995).

In terms of perceptual speed, a meta-analysis of large scale nationally representative samples of adolescents to young adults (approximately 15 to 22 years of age) found that there was a difference in perceptual speed of d = -0.21 to -0.43 (Hedges & Nowell, 1995). These effect sizes are about the same or greater in magnitude compared to the AFOQT's measure (d = 0.15), and in the opposite direction. The Scan subtest in the ATCS selection process had a male-female difference of d = 0.06 (Outtz & Hanges, 2013). This effect size is in the same direction but is slightly smaller than the already minor difference in the AFOQT.

D. Personality Measures. Table 10 provides a summary of the effect size estimates for the Big Five Dimensions as reported in the literature, in addition to any specific personality facets approximate to the measures used in the AFOQT.

Table 10. Female-Male Standardized Differences in the Literature and AFOQT: Personality Measures

Literature Measure	*d*	Female *N*	Male *N*	AFOQT Test	*d*	Female *N*	Male *N*
Anxiety	**-0.40**[a]	500	500	Stress Under Pressure	**-0.47**	10,543	29,520
Compliance	-0.38[a]	500	500	Dominance-Leader	0.02	10,543	29,520
Assertiveness	0.19[a]	500	500	Unassertive	0.08	10,543	29,520
Hyper-Competitive	**0.46**[b]	139	168	Hyper-Competitive	0.33	10,543	29,518
	-0.09[c] or -0.37[d]	-	-				
Teamwork	0.02[e]	64	148	Team Player	-0.08	10,543	29,519
Openness[f]	0.22	-	-				
Conscientiousness	-0.20	-	-				
Extraversion	-0.15	-	-				
Agreeableness	-0.19	-	-				
Neuroticism	**-0.53**	-	-				

Note: Positive effect sizes indicate that males scored higher, and negative effect sizes indicate that females scored higher. Effect sizes meeting the $d = |0.40|$ threshold are in bold.
[a]NEO-PI-R scores in U.S. adults (Costa & McCrae, 1992; Costa et al., 2001).
[b]Single study on university students (Thornton et al., 2011).
[c]Single study data for college non-medalist athletes. Total non-medalist $N = 81$ (Bhardwaj et al., 2018).
[d]Single study data for college medalist athletes. Total medalist $N = 81$ (Bhardwaj et al., 2018).
[e]Single study on MBA students (Farh et al., 2012).
[f]Meta-analysis on the Big Five dimensions in the U.S. Total $N = 2793$ (Schmitt et al., 2008).

A meta-analysis of the Big Five Dimensions across 55 nations (Schmitt, Realo, Voracek, & Allik, 2008) found that overall, women scored higher on Neuroticism (d = -0.40), and slightly higher on Extraversion (d = -0.10), Agreeableness (d = -0.15), and Conscientiousness (d = -0.12). There was no substantial difference in Openness to Experience (d = 0.05). The differences in the U.S. were largely similar, although effect sizes tended to be marginally larger, and there was a small difference in Openness with men scoring higher, instead of no difference. Effect sizes from the U.S. are shown in Table 10.

Another meta-analysis that used the NEO-PI-R to measure personality across countries provided comparisons Stress Under Pressure, Dominance-Leader, and Unassertive (Costa, Terracciano, & McCrae, 2001). For the U.S., An Anxiety measure found a difference of the same direction and magnitude as Stress Under Pressure. A measure of Compliance, as a potential inverse of Dominance-Leader, had a difference with women showing more compliance (and assumedly less dominance), in the U.S. while the AFOQT found no difference. Note that this comparison is rough. While Unassertive in the AFOQT showed

minimal difference between men and women, Assertive in Costa et al. showed a small difference with men in the U.S. being more assertive. All data from the U.S. in this meta-analysis were drawn from one large study (i.e., Costa & McCrae, 1992).

While no meta-analyses have been conducted investigating Hyper-Competitive or Team Player, some studies have investigated male-female differences for these traits. First, a study on university students (N = 307) found that men tended to have higher levels of hyper-competitiveness than women (d = 0.46; Thornton, Ryckman, & Gold, 2011). This finding was not replicated in a study on 162 athletes from 18 to 25 years old, which found that female athletes had higher levels of hyper-competitiveness (d = -0.09 or -0.37; Bhardwaj, Hooda, & Rathee, 2018). The study on university students was more like the difference found in the AFOQT (d = 0.33) in direction and magnitude. Second, a study from full-time professionals in an MBA program (N = 212) found no male-female difference in teamwork effectiveness (d = 0.02) as rated by supervisors (Farh, Seo, & Tesluk, 2012). A similar result was found in the AFOQT, albeit in the opposite direction (d = -0.08).

2.3.5 Aviation Information - All Groups

While little research explores group differences on tests of aviation knowledge, it is helpful to explore relationships between subgroup aviation participation when examining group differences in the Aviation Knowledge subtest. Currently, the FAA does not report demographic data related to ethnicity or race on civilian pilot certification holders. However, we can assume that aviation knowledge and experience may be more common to certain cultures and socioeconomic groups. Additionally, Ison, Herron & Weiland (2016) evaluated the trends in participation by minorities who completed professional pilot education programs in the United States. They evaluated data collected from the Integrated Postsecondary Education Data System. These participation rate data were compared to demographic data found within the aviation industry. They also reviewed trends over a ten-year period and found minority participation increased from 17.1% to 22.2% as well as strong gains among Hispanics, marginal gains by Asians, and minor decreases among women. Overall, they determined that while minority participation has been growing, minority participation in pilot education remains low when compared to the white-male majority group. This lower participation rate implies that majority groups may be more likely to obtain knowledge of aviation principles outside of explicit study for the AFOQT.

3.0 EXTENDED DISCUSSION OF METHODS TO REDUCE ADVERSE IMPACT

3.1 Subtests for which Adverse Impact Exists

A mean group difference at or above 0.40 standard deviations for at least one subgroup was found for every subtest in the AFOQT except for Team Player, Unassertive, Hyper-Competitive, and Dominance-Leader. The subtests that had adverse impact for the most subgroups were Instrument Comprehension and Aviation Information, which both had adverse impact for Black, female, and Asian examinees. Across the subgroups, the largest number of subtests had adverse impact against Black individuals, for whom every achievement subtest had adverse impact, and none of the personality subtests had adverse impact. The size of the Black-White differences for the achievement tests also tended to be larger compared to differences for other subgroups where adverse impact was found. None of the subtests had adverse impact against Hispanic examinees, for whom the largest mean difference was in Verbal Analogies at $d = 0.39$. Asian Americans had the largest differences in favor of the minority group and had many tests for which the mean group differences were very small. However, this group, along with women, tended to have the largest differences on the personality measures with Stress Under Pressure showing adverse impact against both (Table 11).

Methods of reducing subgroup differences were investigated for every subtest for which at least one group met the adverse impact threshold (i.e., $d = 0.40$).

Table 11. Standardized Mean Differences in AFOQT Subtests for Protected Groups

Subtest	Black-White	Female-Male	Asian-White	Hispanic-Non	# of Subgroups with AI
Verbal Analogies	**0.86**	0.19	0.34	0.39	1
Arithmetic Reasoning	**0.93**	**0.54**	-0.10	0.35	2
Word Knowledge	**0.77**	0.24	**0.40**	0.35	2
Math Knowledge	**0.78**	0.39	-0.35	0.30	1
Reading Comprehension	**0.97**	0.37	**0.54**	0.33	2
Physical Science	**0.87**	**0.66**	0.11	0.28	2
Table Reading	**0.82**	0.15	0.10	0.26	1
Instrument Comprehension	**1.15**	**1.08**	**0.40**	0.28	3
Block Counting	**1.03**	**0.49**	0.15	0.22	2
Aviation Information	**0.88**	**0.81**	**0.54**	0.34	3
Team Player (+)	-0.07	-0.08	0.17	-0.06	-
Stress Under Pressure (-)	-0.05	**-0.47**	**-0.40**	0.04	2
Unassertive (-)	-0.09	0.08	-0.30	0.09	-
Hyper-Competitive (-)	0.04	0.33	-0.26	0.04	-
Dominance-Leader (+)	-0.02	0.02	0.18	-0.04	-

Note: AI = adverse impact. Negative effect sizes indicate that minorities/female examinees scored higher, and positive effect sizes indicate that majority/male examinees scored higher. Effect sizes passing the d = |0.40| threshold are in bold. For negatively coded subtests (Stress Under Pressure, Unassertive, Hypercompetitive) negative effect sizes indicate adverse impact against the minority group or women.

3.2 Subgroup Difference Amelioration Methods

3.2.1 Verbal Load

Reducing the reading and language demands of tests has been shown to be effective in reducing adverse impact against racial and ethnic minorities, especially those whose preferred language is not the language used in the test. Much of the evidence for this approach is indirect (i.e., from research on SJTs, tests for children, and ELLs), but promising. The subtests most appropriate for these approaches are the quantitative subtests, along with Table Reading, Instrument Comprehension, Block Counting, Physical Science, and Verbal Analogies.

Substantial research in the area of SJTs has shown the potential for reduced reading requirements. Chan and Schmitt (1997) demonstrated that for parallel written and video SJTs on work habits and interpersonal skills, video SJTs had lower adverse impact against Black participants. Researchers have successfully attributed the reduction in adverse impact of video SJTs and other high-fidelity simulations to their reduced reading or language requirements (Chan & Schmitt, 1997; Lievens & Sackett, 2006; de Meijer, Born, Terlouw, & van der Molen, 2006). Research on individuals learning English as a second language also supports the use of lowered verbal requirements. Researchers have found that linguistic complexity may add measurement error to a test (Abedi, 2006; Abedi, Leon, & Mirocha, 2003). Abedi and Lord (2001) found that removal of unfamiliar or infrequent words and simplified sentence structure on a mathematics test resulted in improvement in test performance for eighth-grade students with poor English ability and poor or moderate mathematics ability. While most groups in the AFOQT did not reach the adverse impact threshold for the quantitative tests (Arithmetic Reasoning and Math Knowledge) many were close (see Table 11), and these results may generalize to other subtests. An important note is that Abedi and Lord (2001) found that lower English ability only acted as a detriment to examinees who were also low or moderate in mathematics ability, so their results would likely only apply to examinees in the AFOQT with low or moderate ability on the test subject. Some linguistic features which may cause undue difficulty for ELLs are long noun phrases; long question phrases; passive voice; comparative structures; prepositional phrases, sentence, and discourse structure; subordinate, conditional and relative clauses; abstract or impersonal presentation and negation; and low word familiarity (Abedi, Lord, & Plummer, 1997)

Research on the reduction of adverse impact through language-reduced tests has also been done for detecting giftedness in minority, English language learners, and low-SES children. Some research suggests that tests that do not rely on reading ability show smaller subgroup differences across ethnicities in the identification of giftedness, but differences can still be substantial (e.g., Naglieri & Ford, 2003; Naglieri & Ronning, 2000; Lewis, DeCamp-Fritson, Ramage, McFarland, & Archwamety, 2007). Analyses for these tests have found other issues, such as a lack of invariance across protected groups (Benson, Kranzler, & Floyd, 2018), and flawed norming groups that cause the number of gifted children to be overestimated (Lohman, Korb, & Lakin, 2008).

For the most part, available evidence suggests validity is not sacrificed for these reductions in adverse impact. Preliminary research on tests without verbal components for children has found good criterion validity for academic success (Naglieri, 2005). A meta-analysis by Christian, Edwards, and Bradley (2010) found that for heterogeneous construct composites, interpersonal skills, and leadership, video SJTs (with low reading requirements) consistently displayed higher criterion-validity than written SJTs (with high reading requirements). The researchers found that certain constructs were more closely related to some criteria than others (i.e., task performance, contextual performance, or managerial performance). However, they cautioned that for some criteria, there were a limited number of studies, which negatively impacted the stability of their results (e.g., teamwork SJTs were found to better predict task performance than contextual performance, which was contrary to their predictions). Note that the validity findings in Christian et al. were calculated for effect sizes collapsed across criterion type.

An important issue regarding the validity of items with lowered language or reading requirements is the possibility that the underlying construct of the test may be changed. Lievens and Sackett (2006) found that video SJTs were more powerful predictors of interpersonal criteria, whereas written SJTs were more predictive of cognitive criteria (although the written SJTs had lower validity overall). Prediction of success in verbal performance areas may also be lower (Lohman & Gambrell, 2012). Physical Science is illustrative for how validity may change when the verbal requirements of the test are changed. Physical Science in the AFOQT is part of the verbal ability factor (Carretta et al., 2016). While reducing the verbal load of the items in this subtest may decrease adverse impact for minorities, the predictive validity of the subtest may be reduced to the degree that reading ability contributes to performance.

There are many considerations when attempting to implement an effective language- or reading-reduced assessment. In the field of giftedness research, language competency and reading ability have proven to not be the only critical factors for subgroup differences. Research has suggested that figural reasoning tests (such as Raven's progressive matrices; Raven, 1938) are not culturally blind even though they have little verbal content (Anastasi & Urbina, 1997; Brouwers, Van de Vijver, & Van Hemert, 2009; Kendall, Verster, & Von Mollendorf, 1988; Lohman, Korb, & Lakin, 2008; Weiss, Saklofske, Prifitera, & Holdnack, 2006). Additionally, some nonverbal tasks may

also increase subgroup differences for certain groups. Black individuals have been found to perform relatively poorly on nonverbal tasks, with differences larger for those that require spatial abilities (Jensen, 1980; Schmitt et al., 1996; Sattler, 2008), a similar finding to gender differences in spatial reasoning (e.g., Maeda & Yoon, 2013). Picture-based nonverbal formats have been proffered as a more effective alternative for predicting academic excellence and reducing adverse impact than figural tests (Lohman & Gambrell, 2012). In this vein, an alternative to Verbal Analogies might be visual analogies. Translating these items to pictographic forms has been shown to be effective in the evaluation of children (see Lohman & Gambrell, 2012). A weakness of this approach is that the pictorial analogy items were produced for young children and it may be difficult to produce such items with sufficient complexity to measure adults.

One method that avoids these issues is to focus on the reading and language requirements of the instructions. Research has suggested that difficulty understanding instructions of unfamiliar tasks might lead to artifactually reduced scores on the test for low-SES, bilingual, and minority children due to the adoption of inappropriate strategies or approaches to solving the items (Budoff, Gimon, & Corman, 1974; Hessels & Hamers, 1993; Resing, Tunteler, de Jong, & Bosma, 2009; Scarr, 1994; see also Lohman & Gambrell, 2012 for a review). This applies to Block Counting, Instrument Comprehension, and Table Reading, as examinees are unlikely to have encountered tasks such as these in their day-to-day experience or schooling. Some examinees rely solely on the instructions to understand the test. Thus, the instructions should be reasonably comprehensive without complex language and supply multiple representative examples. Past research on the AFOQT has noted reduced scores potentially due to instructions. Carretta, Rose, and Trent (2016) found that certain types of items in the Block Counting subtest for form T were disproportionately difficult and conjectured that this was due to the item type not being represented in the sample items.

Another method that is minimally invasive to the item content would be to supplement items in Physical Science and Arithmetic Reasoning with visual diagrams and aids. These subtests rely on written descriptions of situations, so visual descriptions of the situations would help examinees with lower English ability understand the item stems. Qualitative research has also indicated that fifth-grade students from low-income homes and ELLs may interpret science items differently,

regardless of actual science knowledge (Noble et al., 2012). Simplifying the language itself may also be effective, such as removing unfamiliar words and simplifying sentence structure (Abedi & Lord, 2001).

3.2.2 Stereotype Threat

An extensively researched and controversial method is stereotype threat reduction. As discussed below, many methods exist for stereotype threat reduction, but few are applicable to a testing environment, and the efficacy of those that are may be very limited. Caution should be given to certain more extensive interventions, as inappropriate application could lead to increases in subgroup differences. Stereotype threat applies to any test for which some group is generally believed to perform poorly. Because of this, any of the tests in the AFOQT could potentially be impacted.

Stereotype threat is described as the phenomenon under which individuals belonging to a certain subgroup underperform on a test when their group is stereotyped to do poorly in the test's subject (Steele & Aronson, 1995). This underperformance has been theorized to be due to stress individuals feel related to the risk that they will confirm the negative stereotype about their group (as this stress depletes their working memory capacity), loss in motivation due to the stereotype, and other processes (Aronson & McGlone, 2009; Spencer, Logel, & Davies, 2016). One of the most popular methods of experimentally reducing stereotype threat is simply postponing inquiry of demographic information (e.g., race, gender, ethnicity, name) until after the test is administered to avoid making the groups examinees belong to salient to them while they take the test.

Stereotype threat as a subject has garnered a considerable amount of research. Many meta-analyses relevant to testing have been conducted (i.e., Appel, Weber, & Kronberger, 2015; Doyle & Voyer, 2016; Nadler & Clark, 2011; Nguyen & Ryan, 2008; Picho, Rodriguez, & Finnie, 2013; Shewach, Sackett, & Quint, 2019; Stoet & Geary, 2012; Walton & Cohen, 2003; Walton & Spencer, 2009; a reanalysis: Zigerell, 2017). The most recent and largest meta-analysis (i.e., Shewach et al., 2019) separately analyzed lab studies with conditions the researchers believed most closely approximated real testing situations. That is, they excluded studies that controlled for prior ability (e.g., past scores on the SAT), scored performance as total correct divided by the number attempted, or compared inducement of stereotype threat to stereotype threat reduction rather than stereotype

threat inducement to a control group. Sewach et al. also included only studies that subtly induced the negative stereotype (e.g., requesting demographic information before the exam to remind examinees of their group status rather than explicitly telling examinees that certain groups performed poorly on the test). These restrictions produced an effect size of d = -0.14, which was reduced to d = -0.09 when corrections for possible publication bias were made.

Shewach et al. (2019) reported that they excluded studies that compared stereotype inducement to reduction because a common method of reducing stereotype threat (i.e., telling examinees that all groups perform equally) would be considered unethical in real-world testing scenarios and thus could not be implemented. While, as noted by Shewach et al., this manipulation is frequently used to decrease stereotype threat in laboratory experiments, it is not the only method that has been explored. Other research has also reduced stereotype threat by asking participants to list characteristics shared by individuals in both groups (Rosenthal & Crisp, 2006), providing an alternative explanation for test anxiety (i.e., misattribution; Ben-Zeev et al., 2005), having participants write a self-affirming essay (Bowen et al., 2013), having participants select their most important values from a list then explain their choices (Cohen et al., 2006), providing same-group role models (Marx & Roman, 2002; Huguet & Régner, 2007), and treating intelligence as incremental and developable (Good, Aronson, & Inzlicht, 2003). Other methods for reducing stereotype threat can be found in "Stereotype and Social Identity Threat" (Aronson & McGlone, 2009). If such methods of reducing stereotype threat can be used, they may prove more effective in producing score improvements. Shewach et al. found that the effect size for studies that compared conditions where stereotype threat was induced to conditions where stereotype threat was reduced was statistically significantly larger than studies comparing stereotype induction to a control group (d = -0.39 compared to d = -0.28). An older meta-analysis that included only studies with interventions to reduce stereotype threat found that minority individuals outperformed majority individuals with the same level of past performance in both laboratory studies and field studies (Walton & Spencer, 2009).

Despite the optimistic view presented in studies that implement interventions for stereotype threat, the degree to which interventions can impact real testing situations is still uncertain. First, Shewach et al. (2019) considered motivation and found the effect size for studies that provided monetary or

other rewards for high performance was small (i.e., d = 0.00 to -0.14). Shewach et al. argued that situations where performance is attached to rewards should more closely mirror real testing situations. The researchers noted that only a small proportion of studies included factors to increase motivation (11 out of 181 samples). Academic field studies included in the Schewach et al. meta-analysis also cast doubt on the external validity of stereotype threat reduction, as the effect size found here was d = -0.01 (based on four studies with large samples, N = 1,670). Upcoming research regarding stereotype threat will attempt to clarify whether attempts to reduce the effect have an appreciable impact on mean score differences among groups. Lewis and Michalak (2019) submitted a cross-temporal meta-analysis design to resolve mixed findings on the existence of stereotype threat in the literature. A study plan by Forscher et al. (2019) aims to address issues with a small sample size in past literature by recruiting a large sample of Black students. They will also address inconsistencies in operationalization of stereotype threat. Specifically, some researchers have attributed failures to replicate stereotype threat to insufficient manipulation of threat levels (see Spencer et al., 2016). Forscher et al. plan to address this by comparing frequently used operationalizations of stereotype threat induction and reduction.

Potentially, stereotype threat research offers a low-cost and effective method of ameliorating subgroup differences in test performance. However, stereotype threat reduction methods could also have little or no effect on test performance. Some field studies have found that certain operationalizations of stereotype threat reduction decreased the performance of a subgroup (e.g., Gillespie, Converse, & Kriska, 2010, who replaced conventional methods with race-affirming ones), while others have found small (meta-analysis: Walton & Spencer, 2009) or slight (Danacher & Crandall, 2008) improvements. Even if stereotype threat reduction methods are found to impact performance, they may not benefit all groups equally. Stereotype lift, or increased performance for groups stereotyped to do better on the task, may be decreased as well. The loss of stereotype lift could potentially lower the performance of certain groups in certain areas (e.g., Asian examinees on quantitative tests). However, evidence on White men suggests that this effect is small (d = 0.24) and not expunged easily (Walton & Cohen, 2003). Because of the mixed results in the literature on stereotype threat, firm conclusions cannot be drawn. However, the small degree of risk and relative ease of implementation mean that few drawbacks exist for implementing this method.

3.2.3 Adding Low-Impact Tests

A substantial body of literature demonstrates that adding low-impact constructs to test batteries decreases subgroup differences without damaging validity (see Sackett et al., 2001; Ployhart & Holtz, 2008). This method shrinks the subgroup differences for the test as a whole. Below we review low-impact tests found in the literature.

The Sternberg Triarchic Abilities Test (STAT; Sternberg, 1993) has shown promise in reducing adverse impact and increasing the validity of tests for educational performance. A project sponsored by College Board administered versions of STAT tests on creativity, practical skills, and analytical skills along with the SAT to college students at 13 colleges and universities across the United States. The additional subtests increased predictive validity over the SAT and high school GPA. Additionally, the STAT measures were found to lower the overall adverse impact (Sternberg, The Rainbow Project Collaborators, 2006). However, benefits were smaller for Asian examinees. Research on practical skills administered with the Graduate Management Admissions Test indicated that Asian and Hispanic examinees performed worse on the practical measures. This difference was found to be largely due to differences in citizenship status (Hedlund, Wilt, Nebel, Ashford, & Sternberg, 2006). While the samples for several subgroups were small, additional research has also supported the results found by Sternberg and The Rainbow Project Collaborators with the GMAT (Hedlund, et al., 2006), Advanced Placement exams for statistics and psychology (Stemler, Grigorenko, Jarvin, & Sternberg, 2006), and Advanced Placement exams for physics (Stemler, Sternberg, Grigorenko, Jarvin, & Sharpes, 2009). These results across diverse subjects suggest the addition of creativity and practical skills to the AFOQT may increase predictive validity and lower adverse impact for training performance.

Implementing alternative measures of intelligence (e.g., logic, reasoning) was one of the three most highly recommended methods in a review by De Soete et al. (2013). One alternative to traditional methods of predicting success (e.g., verbal reasoning, quantitative reasoning) was information processing. Information processing tests (i.e., tests of individual ability to direct, hold, and control thoughts and similar processes) were found in a recent meta-analysis to be related to work and academic performance and to have lower Black-White subgroup differences than conventional intelligence tests (Larson, 2019). The author of the meta-analysis noted that most

popular intelligence tests operationalize intelligence by the amount of knowledge a person has, whereas the information processing tests measure a person's ability to think. Individuals necessarily vary on the knowledge they have learned, and this can reflect their racial subgroup, exacerbating group differences (Fagan & Holland, 2002; 2007; Malda, van de Vijver, & Temane, 2010). Most of the tests in the AFOQT can be characterized at least partially as tests of knowledge gained. Some alternative intelligence tests, such as the Siena Reasoning Test, have also been found to have higher validity than traditional tests in some cases (Ferreter, Goldstein, Scherbaum, Yusko, & Jun, 2008; Yusko & Goldstein, 2008).

A recent report produced for the Air Force outlined promising alternative intelligence measures that are relatively independent from culture (e.g., Working Memory, Fluid Intelligence, Learning Agility; Teachout, Shore, Martinez, & Wolliston, 2019). Some researchers categorize Working Memory as a subtype of Fluid Intelligence and others do not (Hanges & Feinberg, 2010). While little research has been conducted on the subgroup differences for Working Memory measures, Larson (2019) concluded that the initial research here was promising, as two studies (i.e., Malda et al., 2010; Nelson, 2003) found reduced differences between Black and White examinees. Further, Working Memory was found to produce incremental validity over the ASVAB across Army, Air Force, and Navy technical schools (Wolfe, 1997). Fluid intelligence has been found to produce lower subgroup differences than crystallized intelligence (Hough et al., 2001), and to predict pilot training performance (Kock & Schlechter, 2009). Learning Agility represents the ability to learn from experience and apply new knowledge into novel situations (Lombardo & Eichinger, 2000). Learning Agility is mostly studied for identification and development of high potential employees and leaders. This construct is a promising construct as initial research has found small subgroup differences (De Meuse, Dai, Eichinger, Page, Clark, & Zewdie., 2011; Capretta Raymond, 2006; Lombardo & Eichinger, 2000). Research on Learning Agility has also found support for the validity of the construct (Capretta Raymond, 2006; Church & Desrosiers, 2006). Learning Agility has also been found to be relatively unrelated to intelligence test scores, academic performance, and all the Big Five Dimensions except for Openness to Experience (Connolly, 2001; De Meuse et al., 2011)

In an Air Force-wide survey, training instructors rated Reasoning (measure soon to be available in the Manpower Test Battery) and Problem solving as some of the most important constructs for success (Shore, Peña, Gonzalez, Haight, & Wolliston, 2019). Larson's (2019) meta-analysis also contains an extensive list of alternative intelligence measures and intelligence measures designed to produce smaller subgroup differences.

While the addition of an integrity test has traditionally been found to be effective in reducing subgroup differences while increasing validity, recent research has not always provided support. Earlier research on integrity tests includes a meta-analysis by Schmidt and Hunter (1998) and Ones et al. (1993). Integrity tests were identified by Schmidt and Hunter as having the most incremental validity over cognitive ability tests for predicting job performance. Ones et al. found in a meta-analysis that criterion validity of integrity tests for job performance was .41 and that it was unrelated to cognitive ability. A more recent meta-analysis with stricter inclusion criteria by Van Iddekinge, Roth, Raymark, and Odle-Dusseau (2012) had mixed results. Van Iddekinge et al. found that integrity tests had some validity for involuntary turnover (.19), and a small amount of predictive validity for job performance (.15), but greater validity for counterproductive work behavior (CWB; .32). However, Van Iddekinge et al. also found that the validity for CWBs was moderated by type of criterion and study method such that the validity was larger for self-reported CWBs in incumbent and concurrent studies than other-reported CWBs in applicant and predictive studies (.11 for predictive applicant studies). Validities for job performance were also moderated by author (test publisher or independent researcher). There was no validity when all test publishers were excluded from the analysis (.04). Van Iddekinge et al. noted that although validities were generally higher for test publishers, they could not detect any malpractice in the reports, and thus publisher reports may be more accurate than independent research estimates.

Ones and Viswesvaran (1998) found that overt integrity tests had no substantive race differences (for Black, Hispanic, Asian, and Native American examinees compared to White examinees), and small gender differences in favor of women. Bernardin and Cooke (1993) found no correlations between race or gender with performance on the tests. A more recent study on adverse impact in overt integrity tests found that when test items were grouped into facets, the facets with higher levels of adverse impact also tended to have higher validities (Van Iddekinge, Taylor, & Eidson,

2005). However, the study included only small numbers of minority races and ethnicities. Overall, the evidence that integrity tests have both validity and low adverse impact is weak.

Other concerns apply to integrity testing as well. Research has shown that individuals can learn to fake good on integrity tests, and that overt integrity tests are more susceptible than personality-based tests (Alliger & Dwight, 2000). Concern over how integrity tests are received by applicants has also been expressed, but integrity tests have been found to not produce strong negative reactions in applicants (Berry, Sackett, & Wiemann, 2007). Applicant reactions were also impacted by organizational explanations for the test (Berry et al., 2007). Air Force research has also shown that out of a list of 29 competencies, 'Integrity' was found to be lacking in much fewer AFSCs at the end of training than almost all other competencies (Shore et al., 2019). Integrity was rated as important for success across AFSCs, but it does not appear to be a high priority selection measure for the Air Force.

Motivational constructs may be beneficial in prediction and reduction of subgroup differences for officers. Constructs such as Achievement and Responsibility (facets of Conscientiousness) have demonstrated predictive validity for performance (Teachout et al., 2019). Achievement and Responsibility were rated by Air Force training instructors as some of the most important attributes for success in officer AFSs, and they currently have measures in the TAPAS (Shore et al., 2019). Another attribute rated highly by trainers was Self-Discipline, a measure available from the SDI-O. No relevant research was located regarding the subgroup differences of these constructs. Such a change may particularly benefit women, as some research has shown that higher levels of academic effort/discipline partially explain female underprediction of academic performance from achievement tests (e.g., Keiser, Sackett, Kuncel, & Brothen, 2016; Mattern, Sanchez, & Ndum, 2017). However, small subgroup differences found for Conscientiousness and its facet Achievement suggest that this method should benefit racial and ethnic groups as well (Foldes et al., 2008).

Finally, biodata has shown the potential for reducing adverse impact. Biodata is a method of determining an individual's past experiences, interests, and attitudes, and using these to predict performance in relevant domains. The level of adverse impact generated by biodata is lower than

that of general mental ability tests (d = 0.33, Bobko, Roth, & Potosky, 1999). However, Bobko and Roth (2013) demonstrated that the actual value is likely slightly larger (d = 0.39) due to range restriction in many studies. The degree of adverse impact also appears to vary by protected class. One review of a government biodata instrument measuring educational achievement, work competency, and leadership skills found standardized mean differences of 0.27, 0.08, and -0.15 for Black individuals, Hispanic individuals, and women respectively (Gandy, Dye, & MacLane, 1994). Differences in the size of the effect may also depend on the construct measured. As an example of relevant constructs, a barrier analysis for Air Traffic Control Specialists reported subgroup differences for an instrument assessing work attributes from past experiences (Outtz & Hanges, 2013). The standardized subgroup differences for Asian, Black, Hispanic, and female examinees are reproduced in Table 12. Note that only individuals aiming to become air traffic controllers take this measure.

Table 12. Biodata Cohen's *d* Effect Sizes, Selection for Air Traffic Control Specialists

Construct	Asian-White	Black-White	Hispanic-White	Female-Male
Composure	0.20	-0.04	-0.01	0.11
Consistency of Work Behavior	0.22	-0.10	0.02	-0.03
Concentration	0.13	-0.13	-0.08	-0.04
Decisiveness	0.22	-0.07	-0.01	-0.08
Self-Confidence	0.25	-0.06	-0.08	0.03
Interpersonal Tolerance	0.07	-0.20	-0.10	-0.09
Execution	0.00	-0.01	-0.08	-0.07
Task Closure/Thoroughness	0.15	-0.16	-0.06	-0.25
Flexibility	0.22	-0.10	-0.03	-0.05
Self-Awareness	0.07	0.08	0.07	-0.10
Sustained Attention	0.03	-0.06	-0.02	-0.02

Note. positive effect sizes indicate majorities/men scored higher, negative effect sizes indicate minorities/women scored higher. Effect sizes from "Barrier Analysis of the Air Traffic Control Specialists (ATCS) Centralized Hiring Process" by Outtz, J. L., and Hanges, P. J. (2013). Washington, DC: Outtz and Associates. https://www.faa.gov/

Guidelines for development of biodata instruments demand that they are selected on a rational, theoretical basis instead of a purely empirical one (e.g., Stokes & Cooper, 2001). Processes such as these can prevent the use of items that are highly correlated with certain subgroups or demographic factors. The rational development of biodata items is especially important considering Whitney and Schmitt (1997) found that over a fourth of biodata items they analyzed produced differential item functioning (DIF). Imus et al. (2011) found that the items with DIF in

biodata were identifiable by sensitivity review. Dean (2013) found that the removal of response options for which there was DIF reduced ethnic group differences without harming validity. Thus, biodata may represent a measure with readily controllable subgroup differences.

In addition to relatively low subgroup differences, biodata has shown to have high criterion validity for job performance. Reilly and Chao (1982) found validities of 0.14 to 0.52 depending on occupation, and Schmidt and Hunter (1998) found an overall value of 0.35. Research by Mount, Witt, & Barrick (2000) also demonstrated that biodata accounted for incremental validity over personality and general mental ability. However, biodata is not without issues. Depending on the questions asked, biodata may not have face validity, or it may invade the privacy of examinees. Like any non-cognitive tests, faking is also a concern (McFarland & Ryan, 2000). Research has shown that issues with faking can be ameliorated by ensuring that all items are verifiable (Kluger & Colella, 1993; Harold, McFarland, & Weekley, 2006), and by asking individuals to elaborate on their answers (which reduces self-serving biases in memory; Schmitt & Kunce, 2002; Schmitt, Oswald, Gillespie, & Ramsay, 2003). Another important consideration for biodata is that it should be validated on applicant samples, as research has shown that items valid for incumbents are often not valid for applicants (Stokes, Hogan, & Snell, 1993). A study completed by the Federal Aviation Administration investigated five biodata items (i.e., high school GPA, high school GPA in math, educational degree, Collegiate Training Initiative program graduate or not, pilot certificate or not) for air traffic controllers and found that high school math GPA and holding some type of pilot certificate were significant predictors of training success for en-route centers, but not terminal facilities (Pierce, Broach, Byrne, & Bleckley, 2014).

3.2.4 Golden Rule-Type Adjustments

Golden rule-type methods as discussed here mean that within the appropriate content area items are pre-screened for subgroup differences and test items are chosen beginning with those that produce the smallest differences in addition to psychometric quality. When applied carefully, the method has found support for decreasing group differences without damaging content validity or disadvantaging certain groups (see Kiddler & Rosner, 2002). No empirical research was found regarding criterion-related validity. Use of procedures such as this need to be weighed against other interests. That is, what other criteria will the Air Force use when determining the

acceptability of an item? How highly prioritized will adverse impact be in the item-selection process? As some researchers have commented, Golden Rule-type policies will jeopardize the validity of a test when adverse impact is prioritized above validity and may inhibit detection of biased items (e.g., Geisinger, 1988; Linn & Drasgow, 1987). This process could be applied to all subtests, thereby alleviating subgroup differences in the AFOQT as a whole.

3.2.5 Alternative Measures

One approach to reducing subgroup differences is using an alternate test of the same construct that produces less adverse impact. Our review uncovered such tests for Block Counting and Stress Under Pressure.

First, a valid test with lower adverse impact than Block Counting may be Assembling Objects. Assembling Objects is a spatial visualization test requiring the ability to figure out how an object will look when its parts are put together. While this test is given as part of the ASVAB and not used to make decisions about Air Force officers, it serves as an example of a spatial ability test with small group differences. Overall gender differences in Assembling Objects have been found to be small (e.g., d =0.06; Russell & Peterson, 2001; d = 0.08, Sackett, Eitelberg, & Sellman, 2009). This effect size is much smaller than gender differences for other spatial ability tests. Likewise, research on Assembling Objects indicates that the test reduces adverse impact and score barriers for both women and ethnic/racial minority groups when included in the ASVAB (Held & Carretta, 2013). Anderson et al. (2011) concluded that adding the Assembling Objects test to the AFQT composite score would increase the AFQT's performance and job knowledge prediction for Army occupations and decrease subgroup differences between White and Hispanic examinees. However, they also found doing so would increase differences for female and Black examinees. The Assembling objects tests should be explored as a measure of spatial visualization that may provide a reduction in subgroup differences.

Next, Stress Under Pressure may produce lower subgroup differences for women and Asian individuals if a different test methodology is used. Chinese nationals tend to be motivated to express modesty to a greater extent in surveys (Cai et al., 2011; Johnson & van de Vijver, 2003; Kurman, 2003; Lalwani, Shavitt, & Johnson, 2006; Lalwani, Shrum, & Chiu, 2009). Socially desirable responses may also be different for other cultures, influencing how individuals from

these cultures respond to surveys (Uskul, Oyserman, & Schwarz, 2010). However, these findings may not generalize to Chinese Americans or other Asian Americans (Bachman, O'Malley, & Freedman-Doan, 2010). Women have shown a greater tendency to characterize themselves in terms of stereotypes about women than men characterize themselves with stereotypes about men (Cadinu & Galdi, 2012; Cadinu, Latrofa, & Carnaghi, 2013; Guimond, Chaterd, Martinot, Crisp, & Redersdorff, 2006; Lorenzi-Cioldi, 1991). A DIF analysis found that men and women differed in the types of negative affectivity they endorsed (e.g., men preferentially endorsed irritability and tension while women endorsed emotional vulnerability and sensitivity; Smith & Reise, 1998). These findings regarding cultural and self-stereotyping differences should apply to self-description inventories like Stress Under Pressure.

A potential resolution for these differing tendencies is to evaluate the predictive power and group differences of different operationalizations of stress coping and select the one that is optimal for these measures. Different measures should activate cultural norms that produce differences to varying degrees. A Situational Judgment Test (SJT) may be a viable alternative, as Pangallo, Zibarras, and Patterson (2016) were able to produce a valid and reliable SJT for resiliency, a related construct. The review of constructs performed for the Air Force identified several other potential measures. The most similar to Stress Under Pressure in the review were Defensive Reactivity and Adjustment (Teachout et al., 2019). Adjustment was also determined to be one of the most important attributes for success as an Air Force officer. It has an available measure in the TAPAS and SDI-O (Shore et al., 2019). Different measures of stress-related constructs may have larger or smaller differences depending on the subgroup. For example, different facets of Neuroticism display widely disparate group differences depending on the subgroup. In the Foldes et al. (2008) meta-analysis, Low Anxiety showed a small minority advantage for Asian and Hispanic groups (at d = -0.27 and -0.25 respectively) but a comparable disadvantage for Black individuals (d = 0.23). On the other hand, Even-Tempered showed a disadvantage for Asian individuals at d = 0.38, but no substantial majority-minority difference for Hispanic or Black individuals (Foldes et al., 2008). The dimension Neuroticism did not pass d = |0.12| for any of the racial or minority groups (Foldes et al., 2008), but was d = -0.53 for female examinees (Schmitt et al., 2008). Because of this pattern, using another facet or inventory to measure stress may benefit some subgroups but

harm others. Changes to predictive validity may also occur, given that different facets measure related, but distinct, traits (e.g., Hough & Ones, 2002).

3.2.6 Structured Interviews

Structured interviews may serve as a subgroup difference reduction method but require too high of a cost, personnel, and time commitment to be used on a large scale. Structured interviews are interviews that standardize the interview questions, observations, and ratings (Levashina, Hartwell, Morgeson, & Campion, 2014). Extensive evidence has shown that structured interviews both have good validity (Campion, Palmer, & Campion, 1997; Levashina et al., 2014) and relatively small subgroup differences for minority race, ethnicity (Bobko, Roth, & Potosky, 1999; Huffcut & Roth, 1998), and gender (Alonso, Moscoso, & Salgado, 2017) groups. Note however that evidence for ethnic groups found that correction for range restriction may lead to substantially larger effect sizes (Roth, Van Iddekinge, Huffcutt, Eidson, & Bobko, 2002). Structured interviews have also shown incremental validity over Conscientiousness and general mental ability measures (Cortina, Goldstein, Payne, Davison, & Gilliland, 2000), although they were found to not have incremental validity for pilot selection (Walters, Miller, & Ree, 1993). However, despite the advantages of structuring an interview, structured interviews also have drawbacks. Structured interviews may be seen less favorably by interviewers and interviewees than unstructured interviews, but research is mixed (Levashina et al., 2014). Other issues with interviews are that they are labor intensive compared to other selection methods (e.g., Ryan & Tippins, 2004). In most cases, interviews must be conducted with a ratio of one applicant to one administrator. Additionally, to minimize subjectivity and standardize interviews, interviewers need to be trained to rate applicants in a standardized way (e.g., Woehr & Huffcut, 1994), which further increases costs and time commitments. Interviews may also be more susceptible to the individual biases of the rater. Many issues may influence the ratings of interviewers, such as the contrast effect, where an interviewee is rated more highly if interviewed immediately after a notably poor interviewee (Wexley, Sanders, & Yukl, 1973). The opportunity for individual biases to enter ratings may threaten the ability of interviews to decrease subgroup differences. Finally, interviewees may also distort their answers to appear more favorable (i.e., impression management), although structured interviews are less susceptible to this (Levashina et al., 2014).

3.2.7 Impacts of Differing Item Answering Strategies

Evidence suggests that gender differences on spatial ability tests may be reduced if slightly more time is provided per item. No definitive evidence exists explaining this phenomenon, but one possible explanation is differences in strategies used to answer items. No comparative research was located for racial or ethnic minorities. The two spatially oriented subtests, Block Counting, and Instrument Comprehension should show a reduction in adverse impact if slightly more administration time is given. Risks exist in terms of how this change may impact the tests' validity, difficulty, and discrimination.

Maeda and Yoon (2013) demonstrated via meta-analysis that a factor explaining variance due to gender in mental rotation scores was administration time. When no time limit was set, the difference was $g = 0.57$; when a time limit of 40, 50, or 60 seconds per item was given, $g = 0.31$; and when time was limited to 18, 24, or 30 seconds per item, $g = 0.67$.[1] Note that Maeda and Yoon did not see a reduction in effect size when the participants had unlimited time to complete items. Maeda and Yoon suggested that that the impact of time restrictions may have been due in part to construct irrelevant variance introduced by "speededness." That is, anxiety may arise from perceived time limits and the speed at which men and women mentally process responses to items. The researchers cautioned that they were unable to find research on which of these time limits best represented the true construct. Maeda and Yoon cite Guay's (1980) recommendation that each item should be allotted 40 seconds to minimize examinees' use of analytic strategies instead of mental rotation, although they note that little research support exists for this time limit.

Based on Maeda and Yoon's (2013) results, slightly increasing time allotted to items appears to be a promising approach. This method is not without risks, however. Increasing the allotted test time may produce a ceiling effect where scores accumulate at or near the highest score. Statistics from the Maeda and Yoon meta-analysis were inconclusive in this regard. Changes in the ability to discriminate among examinees may also be influenced if more time is given on the subtests. Finally, allotting additional time may not be effective for non-spatial subtests. Bridgeman, Trapani, and Curley (2004) increased the time allowed to answer items on the SAT verbal and math sections

[1] Hedge's *g* is an effect size estimate similar to Cohen's *d,* but more accurate when small samples are used.

but found no impact on ethnic, racial, or gender differences. In a review, Sackett et al. (2001) found that increased time frequently increased subgroup differences on educational achievement tests.

Other research investigating extraneous factors that may contribute to the gender difference in spatial ability assessments has led to strategy choice as a possible explanation and training as a potential solution (e.g., Bosco et al., 2004). Maeda and Yoon's (2013) findings may also be explained by men and women choosing strategies with different time requirements, although the researchers themselves did not attribute the variance by time limit to be due to strategic approach. For instance, research on mental rotation has provided evidence that differences in strategies partially account for female-male differences in time taken to complete items. Jorden et al. (2002) conducted an fMRI study that showed men were more likely to rotate the entire object, whereas women were more likely to compare specific features of the object across options (which takes longer). Other research on fMRI has shown that the brain areas active while solving spatial rotation tasks differ between men and women (Hugdahl, Thomsen, & Ersland, 2006; Jordan, Wustenberg, Heinze, Peters, & Jäncke, 2002; Weiss et al., 2003). Some evidence has shown that wholistic (e.g., rotate the whole block) or specific strategies (e.g., rotate block sections) are individual characteristics that generalize to other spatial tasks (Janssen & Geiser, 2010). Note though that people frequently choose different strategies based on task type and difficulty, and not all research has found gender differences in strategy selection (Glück & Fitting, 2003). Hirnstein, Bayer, & Hausmann (2009) found that mean differences between men and women on the mental rotation test (MRT) were reduced when they experimentally controlled for whether all response options were investigated. That is, when they required all respondents to check each response option by including a variable number of options that were correct, the size of the gender difference was reduced (d = 0.76 versus 0.95; note that the total N was 34 evenly split between men and women, and significant effects were found). Other research found that combined training in both analytic strategies (preferred by women) and imagistic strategies (preferred by men) removed the gender difference in chemistry achievement regarding molecule rotation (Stieff, Dixon, Ryu, Kumi, & Hegarty, 2013).

Finally, research has shown the types of spatial skills used to answer items may differ among individuals for spatial orientation. Spatial orientation is a spatial sub-ability proposed to measure

an ability to imagine the appearance of objects from different orientations (i.e., perspectives) of the observer (McGee, 1979). A commonly used instrument for this ability is the Guilford and Zimmerman Spatial Orientations Test, in which observers are shown two different views of a landscape from the prow of a boat and have to determine how the boat has changed position from the first to the second view. Research suggests that the Spatial Orientation Test can be completed by using multiple mental strategies. In fact, it has been shown that this test is most commonly solved by using mental rotation strategies (Carpenter & Just, 1986; Carroll, 1993). While there has been some debate about the distinction between spatial orientation and spatial visualization abilities, Kozhevnokov & Hegarty (2001) provided evidence that the two spatial abilities were distinct through the development of a purer measure of spatial orientation. These findings indicate that care should be taken that an instrument measures the intended construct, as different spatial abilities may be able to contribute to test performance.

This section suggests that training respondents on the most efficient item-answering method may improve female scores on spatial tests. Another section of this report (i.e., 3.2.11 Adjusting for Differential Test Exposure) considers coaching and training for all subtests, with less promising results. For spatial tests, careful consideration must be given to the strategies that are trained, as evidence suggests that the strategies most effective for women may not be the same as those effective for men (Stieff et al., 2013). A meta-analysis found increases in spatial skill could be produced by training (g = 0.47), and that available evidence suggested these changes to be reasonably durable (Uttal et al., 2013). The researchers also found that increases from training one spatial skill could transfer to other spatial tasks with sufficient effort. These results indicate that training spatial skills may be helpful as well.

3.2.8 Adjusting for Differential Test Exposure

One area with substantial research is the use of coaching or other test training and retesting as a method of reducing subgroup differences (see De Soete et al., 2013; Sackett et al., 2001; Ployhart & Holtz, 2008). The rationale behind this approach is that ethnic groups differ in their test familiarity and testing skills (De Soete et al., 2013; Sackett et al., 2001). For instance, previous research has suggested that Black examinees tend to utilize ineffective test-taking strategies (e.g., choosing randomly when guessing, choosing the longest response option) more than White

examinees (Ellis & Ryan, 2003; Dollinger & Clark, 2012). Overall, this approach method has produced inconsistent results in terms of ameliorating subgroup differences. A meta-analysis on coaching, practice, and retesting effects in cognitive ability tests suggested that test-taking abilities appear to have small impacts on test performance (Hausknecht, Halpert, Di Paolo, & Gerrard, 2007). Evidence suggests that Block Counting and Instrument Comprehension are the AFOQT subtests that may benefit the most, regardless of the overall effects.

For retesting, score improvements for retesting appear to vary by subgroup. Women and younger individuals have been found to improve their scores at retesting more than men and older individuals (Schleicher, Van Iddekinge, Morgeson, & Campion, 2010; Van Iddekinge, Morgeson, Schleicher, & Campion, 2011). Research has also found that minority race and ethnicity examinees may improve less at retest (Van Iddekinge et al., 2011) although this may vary by test type. Schleicher et al. (2010) found that Black, Hispanic, and Asian examinees tended to improve less than White examinees on written tests, but more or not differently on interviews. However, other research has failed to find moderation of retest improvements based on race or gender (Randall, 2012; Randall, Villado, & Zimmer, 2016). An important note is that minority groups may still benefit from retesting despite not benefiting as much as other groups if the additional attempts allow a greater number of minority examinees to achieve qualifying scores (see Randall, 2012; Randall & Villado, 2017; Randall et al., 2016). More promising results have been found for validity.

While some researchers have expressed concern that retesting may contaminate the test with irrelevant constructs (e.g., Randall & Villado, 2017), criterion-related validity appears not to suffer at retest (Villado, Randall, & Zimmer, 2016) and may even improve (Van Iddekinge et al., 2011). In accordance with these findings, Reeve and Lam (2007) and others (Olenick, Bhatia, & Ryan, 2016) found that size of score increase due to retesting was negatively related to the g saturation of the scale. They also found that gains were positively related to beliefs about testing and motivation for certain scales (Reeve & Lam, 2007).

Other moderators in regard to retesting have also been found. In general, greater score gains have been found on tests with heterogenous item types (Villado et al., 2016), and for individuals with

moderate scores (Randall, 2012). Additionally, a meta-analysis demonstrated that there appear to be diminishing returns such that improvement on the third administration is significantly smaller than improvement on the second administration, and no further gain in scores is apparent after the third attempt (Scharfen, Peters, & Holling, 2018). This research suggests that one additional testing opportunity in the AFOQT over the two currently allowed may benefit examinees, but additional retesting opportunities after that would be unnecessary. However, Scharfen, Peters, & Holling (2018) noted that retest gains on the third attempt were significantly higher for numerical tests than other types of tests (i.e., mixed, verbal), and there were no available studies of fourth administrations for numerical tests.

For coaching, results have shown consistent small score gains for all groups, and inconsistent results as to whether minority or majority groups improve to a greater extent (Sackett et al., 2001). Stieff et al. (2013) found that training chemistry students with a combination of strategies reduced score differences between men and women. An important consideration for their study was that different strategy training approaches benefitted men and women differently (Stieff et al., 2013). The results of Stieff et al. indicate that the AFOQT subtests for which coaching may be effective are the spatial tests, Block Counting, and Instrument Comprehension. Research has indicated that spatial ability is in part related to exposure to tasks requiring mental understanding of relations within space, such as sports or video games (Moreau, Clerc, Mansy-Dannay, & Guerrien, 2012; Feng, Spence, & Pratt, 2007). Research also indicates that individual levels of spatial ability increase with training (Uttal et al., 2013). Men and women improve equally in spatial tasks when trained without special attention to gender-preferred strategies (Uttal et al, 2013), meaning that while subgroup differences may not decrease, more women may achieve qualifying scores. Although difficult, transfer of spatial skills to other spatial tasks appears to be possible with extensive training, and limited evidence suggests changes to be durable (Uttal et al., 2013). Therefore, training spatial skills may generalize to better performance in spatial tasks during job training or on the job.

3.2.9 Item and Response Types

Research on Situational Judgment Tests (SJTs) has shown that particular aspects of how the test is given impact the size of subgroup differences. While some of these adjustments may apply to

other tests in the AFOQT, the most direct application is on the SJT currently administered in the AFOQT.

SJTs have low levels of adverse impact and good validity (Lievens, Peeters, & Schollaert, 2008). SJTs present job candidates with hypothetical scenarios tailored to the job role. Generally, the scenarios are followed by several possible responses where the candidates must choose the most appropriate response. The scenarios and response options are based on a collection of critical incidents, usually gathered through a job analysis. The general features of an SJT, however, may vary in levels of fidelity, response instructions, and scoring methods (Arthur, Glaze, Jarret, White, Schurig, & Taylor, 2014). These factors influence the size of subgroup differences.

First, in terms of fidelity, higher fidelity SJTs have been shown to result in lower subgroup differences. Fidelity is the extent to which a measurement mirrors the true performance environment. Recent research has provided promising results regarding the use of multimedia in SJTs to improve fidelity. For example, video SJTs have shown higher validity coefficients, lower cognitive load, and stronger relationships to interpersonal criteria and leadership skills than paper and pencil SJTs (Christian, Edwards, & Bradley, 2010; Lievens & Sackett, 2006; Lievens & Sackett, 2012). Higher fidelity responses (e.g., acting out a response over writing one over choosing a multiple-choice option) have also been shown to have lower subgroup differences (Lievens, Sackett, Dahlke, Oostrom, & De Soete, 2019) while predicting performance the same or better (Funke & Schuler, 2002; Lievens et al., 2019; Lievens, De Corte, & Westerveld, 2015). Unfortunately, a high-fidelity SJT would demand an online or video format, which is more expensive to develop and administer (Weekly, Hawkes, Guenole & Ployhart, 2015). Investigations into the causes of the lower subgroup differences may shed light on how to produce similar results with less expense.

Several explanations have been offered for the decrease in adverse impact that occurs when fidelity is increased. A highly supported explanation is the decreased reading requirements and thus lower cognitive load (Chan & Schmitt, 1997; Lievens & Sackett, 2006; de Meijer, Born, Terlouw, & van der Molen, 2006). In general, research has shown that measures with more cognitive content have larger ethnic group differences (e.g., Goldstein, Yusko, & Nicolopoulos, 2001; Roth et al., 2008;

Whetzel et al., 2008). This explanation suggests that lowering cognitive or reading requirements may aid in ameliorating adverse impact as well.

Test perceptions have also been investigated as an explanation for the relationship between SJT fidelity and subgroup differences. Chan and Schmitt (1997) found that the face validity of a video SJT was significantly higher than that of a written SJT. A later study by Lievens and Sackett (2006) did not find differences in perceptions of face validity between high and low fidelity SJTs in a high stakes setting. However, their sample was composed of 99.5% White candidates, and the study by Chan and Schmidt found that White examinees saw almost no differences in face validity between the SJT formats compared to Black participants. Edwards and Arthur (2007) also found that differences on a knowledge test were partly mediated by racial differences in the perceived fairness of the test (i.e., a constructed response test was seen as more fair than multiple choice by Black participants), but not its face validity. A literature review by De Soete, Lievens, and Druart (2013) concluded that test perceptions in general had small but positive impacts on subgroup differences. Thus, an alternate route to a high-fidelity SJT may be improving perceptions of the AFOQT in general through providing information on the tests' relevancy for performance (Truxillo, Bauer, Campion, & Paronto, 2002).

Next, response instructions have also been found to influence adverse impact. Typically, SJT response instructions fall into two categories: behavioral and knowledge based. Behavioral tendency instructions ask the respondent to choose the answer(s) or respond based on what their typical response would be. Examples of behavioral tendency instructions include inquiring what the examinee would do in the situation, having the examinee choose the response they would be most and/or least likely to do, to rate and/or rank how likely they would be to enact each response option and to rate their tendency to perform a response option. Knowledge-based instructions ask the respondent to choose the correct or most appropriate/inappropriate response(s). Examples of this approach include: inquiring what an examinee should do in the situation, have the examinee choose the best and/or worst response, have the examinee rank the options from best to worst, and to ask the examinee to rate the effectiveness of each response (McDaniel, Hartman, Whetzel, & Grubb, 2007). Behavioral instructions have shown lower subgroup differences than knowledge instructions. This reduction in adverse impact has been attributed to different loadings on

constructs, as behavioral responses have shown stronger relationships to personality constructs (Whetzel, McDaniel, & Nguyen, 2008). However, faking is a greater concern for behavior-based items.

Finally, the methods chosen for scoring SJTs influence the size of subgroup differences. Arthur, et al. (2014) compared the subgroup differences in SJT performance among African American, Hispanic, Asian, and female groups with different response formats (rate, rank and most/least). The most/least and rank response options demonstrated higher subgroup differences and higher correlation with cognitive ability than the rate response format. McDaniel, Psotka, Legree, Yost, & Weekley (2011) developed methods of scoring tests that reduced Black-White differences and increased validity by correcting for response styles (e.g., tendency to choose extreme or centralized options in a Likert-type scale), which differed between the two groups. McDaniel et al noted that a common method of determining correct answers to SJT items (i.e., using the mean response from experts) rarely resulted in the extreme ends of a Likert-type scale being the correct options, which disadvantages individuals who tend to choose extreme options.

However, these three factors also impact validity, likely through differences in the underlying constructs of the test. Because of this, changing these aspects of the test may produce an SJT that is still valid, but for a different criterion.

First, the fidelity of the SJT influences not only adverse impact but also the SJT's criterion validity. Lievens and Sackett (2006) found that video SJTs were moderate predictors of interpersonal criteria ($r = .35$) but poor predictors of cognitive criteria ($r = .10$). Because of this phenomenon, careful consideration should be given to the important performance criteria before fidelity or cognitive load is manipulated in an attempt to remedy adverse impact.

Next, response instructions also influence the criteria for which SJTs are good predictors. In the McDaniel, Hartman, Whetzel, & Grubb (2007) meta-analysis, behavioral tendency instructions had stronger validities for Agreeableness, Conscientiousness, and Emotional Stability. Knowledge-based instructions had stronger validity for Extraversion, Openness to Experience, and cognitive ability. However, in the meta-analysis, knowledge and behavioral instructions had the

same validity coefficient for job performance (note that the N for behavioral was much smaller than N for knowledge). Nonetheless, is it important to note that knowledge-based instruction has been shown to have higher validity than behavioral tendency responses (McDaniel, Hartman, Whetzel, & Grubb, 2007). It is stipulated that knowledge-based responses assess maximal performance while behavioral-based responses assess typical performance (McDaniel, et al., 2007). These results emphasize again that the desired criterion needs to be considered.

Finally, scoring methods also influence the underlying construct of the SJT, which may impact the relevant criteria. Arthur et al. (2014) found that the correlation between cognitive ability and an SJT varied according to the response format (rank and most/least likely responses had a stronger correlation to *g* than rate responses). However, the effectiveness and effects of the different SJTs design features have not been fully explored (Arthur et al., 2014).

3.2.10 Response Styles

Research has shown that cultures differ in their response styles to Likert-type items, such that some cultures tend to answer more extremely than others (Harzing, Brown, Köster, & Zhao, 2012; Johnson, Kulesa, Cho, & Shavitt, 2005). Black individuals specifically have been found to have a more extreme response style than White individuals (Bachman & Omalley, 1984; Batchelor & Miao, 2016). Promising research was conducted by McDaniel, Psotka, Legree, Yost, and Weekley (2011). They developed scoring methods to decrease Black-White differences in an SJT with a Likert scale while increasing validity. These methods corrected for scatter and elevation. Thus far, these methods have been closely related to SJTs. McDaniel et al. noted that a common method of keying Likert-type items in SJTs (i.e., the average response of experts is the correct response), makes extreme responses rarely correct responses. However, McDaniel et al. argued that their methods should decrease group differences both for groups who avoid and favor extreme responses. Because of this, the method should apply specifically to tests where a certain response style puts an examinee at a disadvantage. Thus, the method may apply to Stress Under Pressure if a certain response style leads to lower scores. While the AFOQT found no differences between Black and White individuals on any of the personality items, there were substantial differences for the Asian and female subgroups for Stress Under Pressure. Note that McDaniel et al. considered

their research to be suggestive, as replications on applicant and cross-cultural samples would need to be done. McDaniel and Weekley (2012) replicated the validity findings of the original study but did not have the necessary sample to re-investigate findings on subgroup differences or to do a cross-cultural replication. Note also that the total validity of the SJT in this replication was low, even though it did increase when the adjustments were made.

3.2.11 Item Content Changes

For Reading Comprehension items, the content of the passages may impact the size of subgroup differences. Evidence has shown that groups perform better on passages with content that is culturally relevant to them. Thus, a method for reducing ethnic and racial subgroup differences in reading comprehension could be to include passages with culturally relevant content. However, while this approach has shown some positive results, questions remain about how culturally specific these items need to be to show an effect.

Keller, Deneen, and Magallan (1991) proposed adding items or passages that relate to a specific group's culture or environment to reduce differential responding. They labeled these types of items as *special interest items.* Research in the area of cognitive psychology has found that understanding the context of a passage impacts comprehension and memory for it (Bransford & Johnson, 1973) and that the perspective of the reader impacts what information is best remembered (Anderson & Pichert, 1978). Examinees with different cultural backgrounds likely differ in their knowledge of the context of passages and view passages on certain topics from different perspectives. Thus, both the passages and the items drawn from passages may be easier or harder depending on the examinee's background. Research has also found that Hispanic and Black examinees perform superiorly (compared to matched-ability White examinees) on sentence completion and reading passages with content especially relevant to Hispanic and Black examinees (Carlton & Harris, 1992; O'Neill & McPeek, 1993; Schmitt & Dorans, 1988). Research on students learning English found that passages with content familiar to the students were easier for them (Alptekin, 2006; Keshavarz, Atai, & Ahmadi, 2007; Yuet & Chan, 2003), although research has been mixed as to whether there is an upper ability limit to this effect (see Yuet & Chan, 2003; Ridgway, 1997). Thus, research on ELL may not generalize to individuals with higher English language skills. Findings have also been uncovered regarding item content for quantitative items.

Research into quantitative reasoning has found that items produce different levels of adverse impact depending on the approach that must be taken to solve them correctly. Generally, items that require "by-the-book" response strategies or algorithms have smaller gender differences than items requiring "out-of-the-box" thinking or unusual items. These differences have been attributed to general tendencies of men and women to approach items in different ways. Gender differences on quantitative items (Math Knowledge and Arithmetic Reasoning) can be ameliorated by removing unusual items, such as described below.

Research on the SAT-M conducted by College Board found that for high-scoring individuals, most items where the average man performed better involved the use of unusual items (i.e., ill-defined items that cannot be solved with regular algorithms), whereas the average woman performed better on usual items (i.e., well-defined items solvable by applying an algorithm; Gallagher, 1992). Similar results were found by Gallagher et al. (2000), whose results indicated that men were more able to appropriately apply different strategies to items (i.e., higher strategy flexibility). They also found that the male advantage was larger on items that required spatial skills, shortcuts, or involving multiple solution paths, than on those requiring verbal skills or school-taught content. The sample size for this study was small (N = 28), which limits the generalizability of the results. Finally, Gallagher, Levin, and Cahalan (2002) found that for the GRE, larger male-female differences were found for items that required spatial abilities rather than verbal abilities, an unconventional solution rather than conventional, and multiple solution paths rather than many steps. However, whether these differences in effect size were significant varied depending on whether the sample was highly skilled in mathematics or not. Gallagher et al. (2002) used these findings to produce a test that had a $d = 0.51$ difference in favor of women, by cloning items that were found to have low subgroup differences (when the researchers included GRE quantitative score as a covariate, the difference between men and women was statistically insignificant at $p < .07$). Two studies have shown that mental rotation ability mediates gender differences in mathematics performance on the SAT (Casey, Nuttall, Pezaris, & Benbow, 1995; Casey, Nuttall, & Pezaris, 1997), so using items that can be solved equally well using spatial or verbal methods could also be an effective approach.

The findings that different types of items produce different gender gaps in performance were mirrored in an international meta-analysis (Else-Quest et al., 2010), which reported different effect sizes depending on the subtype of mathematical ability, with the largest gender difference in the U.S. for measurement (d = 0.15), and the smallest for algebra (d = -0.01). Overall, careful selection and analysis of quantitative items has been demonstrated to impact differences in men and women on quantitative items.

Despite the apparent effectiveness for reducing gender differences, no investigation was found for this approach in terms of validity. Gallagher et al. (2002) expressed that the differences they found may or may not have been relevant to the construct and recommended an evaluation of construct relevancy for the approaches they identified. Whenever item content must be manipulated, changes in the difficulty of the overall test is also a concern. The prototype assessment Gallagher et al (2002) produced by cloning low-impact items had mean percent correct for men at 69.2% and women at 74.8% in a sample of 60 technical science students with high GPAs (A or B students). This average was similar to the performance of technical science students on an operational test form, at 77% to 80% for men and 69% to 77% for women (N = 24,962).

3.2.12 Constructed Response

Constructed response items could be useful for most AFOQT subtests, although there are drawbacks to this method, and evidence is mixed. Arthur, Edwards, and Barret (2002) proposed the use of constructed-response tests as an alternative to multiple choice tests. Constructed response tests require the test-taker to produce the answers to the test rather than choose them (Arthur, Edwards & Barret, 2002). Examples of constructed response tests include fill in the blanks, sentence completion, and short answer tests. Snow (1993) provided a taxonomy of categories of test formats that would fall into the constructed response format. These include simple completion, in which the respondents inserts a word or a phrase to complete a passage; short answer essay, in which the respondents produce a sentence or short paragraph to answer the question; problem exercise, in which the examinees produce and explain a solution to the question; teach-back procedure, in which the examinees explain concepts, procedures, or systems; and long essay, in which respondents need to produce essays or demonstrations.

Arthur, Edwards, and Barret (2002) hypothesized that constructed response items may reduce subgroup differences by decreasing non-job relevant variance (unnecessarily high reading levels), having lesser susceptibility to testwiseness, and having higher face validity. However, subgroup differences in constructed response tests are seldom studied (Lievens, Sackett, Dahlke, Oostrom & De Soete, 2018). Optimistic results regarding subgroup differences for constructed responses have been found in SJT research (i.e., Lievens et al., 2019), although it is uncertain the extent to which results with SJTs would generalize to achievement tests. Available evidence for achievement tests is mixed and suggests that the groups that benefit from constructed response items may depend on the subject. Research on college students found that Black-White differences in a combined math and science test were smaller when constructed response items were used (Edwards & Arthur, 2007), but research on grade school students showed that constructed math items had a similar Black disadvantage to multiple-choice items (Kevelson, 2019). Note that the research on college students had a higher degree of experimental control, but a smaller sample. Results for the math items alone were not reported in the study on college students (i.e., Edwards & Arthur, 2007), so a direct comparison of math items could not be made between the two studies. Evidence from grade school found that constructed response items had a slightly smaller Black-White and Hispanic-White difference than multiple choice for reading comprehension (Kevelson, 2019). No perceptible difference was found between constructed and multiple-choice math items for Hispanic students (Kevelson, 2019). Research on gender differences in math for grade school students found that while there was some variance across grades, in general, there was a larger male advantage on constructed than multiple-choice items (Wilson & Zhang, 1998). Dossey, Mullis, and Jones (1992) believed that constructed format items (i.e., short answer and explain response) in a mathematics test were harder for students to answer but could provide more information on the students' level of proficiency. This assertion is corroborated in the research, which found constructed items to be more challenging (Edwards & Arthur, 2007; Kevelson, 2019).

Little evidence is available regarding the impact of changing multiple choice items to constructed response items has on validity. The available evidence suggests that criterion validity is not impaired (Edwards & Arthur, 2007), but that construct validity is susceptible to change (Rodriguez, 2003).

Constructed response tests have methodological drawbacks; constructed-response tests are less reliable (Rodriguez, 2003) and more expensive to administer and score compared to multiple-choice items, "…the scoring of write-in and mark-in items is more labor intensive and obviously introduces an element of subjectivity in the scoring process" (Arthur et al., 2002, p.1003). Advancements in computer scoring may improve the financial feasibility of using constructed responses (Campion, Campion, Campion, & Reider, 2016).

4.0 OTHER DISCUSSION CONTENT

4.1 Limitations and Future Research

An important limitation of the effect size comparisons between the AFOQT and similar tests in the literature is that this comparison does not indicate the cause of either disparities or similarities in effect sizes between the AFOQT and measures in the literature. Section 2.2 reviews some of the potential explanations. Another consideration for these comparisons is that by necessity, all comparisons were imperfect. Future research should investigate additional comparisons, where relevant influential variables are controlled (e.g., differential range restriction, age, education level, socioeconomic status). Research specifically aimed towards investigations of the relationship between socioeconomic status and score may provide useful insight regarding the influencing factors on AFOQT performance. Future research should also focus on subsets of each demographic group taking the AFOQT. For instance, subtests included in the Pilot composite are taken by all officer applicants, despite that not all officer applicants desire pilot classifications. Motivation effects for these tests likely influence the size of subgroup differences, and thus a clearer interpretation may be made when analyses include just the applicants aiming for pilot classifications. One simple indication of this effect may be the number of examinees in each subgroup who score no points on a subtest. If the proportion is higher in certain groups, motivation differences have likely impacted subgroup differences.

A similar weakness in the review of methods to reduce adverse impact is that some subgroups for which adverse impact was found in the AFOQT did not overlap well with the subgroups included in studies. No adverse impact was found against Hispanic examinees in the AFOQT, but this group represents one of the most frequently researched in the literature. Adverse impact was found

against Asian examinees for several subtests in the AFOQT, but little adverse impact research has been done regarding Asian examinees. Methods of reducing adverse impact for spatial tests focused almost exclusively on women. Because of this disparity, we sometimes relied on assumptions that different subgroups are affected by adverse impact in the same way, which may not be true.

Research should also be done on any amelioration methods that the Air Force considers implementing. While efforts were made to locate relevant validity data and research for situations similar to the Air Force, at times this research simply did not exist. Even when such research was available, there is no guarantee that findings elsewhere will generalize to the AFOQT. Therefore, both subgroup differences across groups and validity should be assessed for any new methods considered.

4.2 Reduction Methods Outside Our Scope

Mean subgroup differences can be influenced by recruitment and sampling strategies, neither of which are addressed in this report. Research has shown that actively recruiting minority individuals with the needed skillset can offset adverse impact (Newman & Lyon, 2009). Additionally, adverse impact can be targeted through post-hoc adjustments as well, such as statistical adjustments. Some researchers maintain that adverse impact as a concept is distinct from subgroup differences, and thus while subgroup differences are reduced using pre-test adjustments (such as the testing methods described here), adverse impact is removed using post-test methods (such as weighting and banding; Arthur, Doverspike, Barrett, & Miguel, 2013).

High fidelity simulations, such as assessment centers, were also excluded from the current review. Despite being widely considered as potentially effective methods of decreasing adverse impact (e.g., De Soete et al., 2012, 2013; Ployhart & Holtz, 2008), high-fidelity simulations were excluded because they constitute both a substantially different selection process instead of a difference in testing methods, and implementation for officer recruits is improbable due to costs and time commitments.

Differential item functioning (DIF) analyses control for the underlying latent variable a measure intends to capture, then compare the frequency of response options across groups. DIF analysis is an extensively researched approach of lowering subgroup differences on tests, but previous research found that DIF was not an issue requiring remedy for the AFOQT Form T (Shore, 2014). The evidence suggests that DIF analyses would have very little or no effect on adverse impact.

FURTHER READING

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

Recommendations:

- Perform job analyses
- Use both cognitive and non-cognitive measures
- Use alternative measurement methods when feasible
    - Interview, SJTs, assessment centers
- Reduce cognitive load, verbal and reading requirements as job analysis permits
- Enhance applicant reactions*
- Consider banding*

*Produces only small impact reductions

Caveats

- Researchers mostly know about Black-White and female-male differences
- Costs must be weighed against the payoff for different methods considered
- A method that reduces impact for one group may exacerbate it for another
- Observed differences are impacted by methodology (i.e., range restriction, reliability)

De Soete, B., Lievens, F., & Druart, C. (2012). An update on the diversity-validity dilemma in personnel selection: A review. *Psychological Topics, 21*(3), 399-424.

Similar but extended content to De Soete, Lievens, and Druart (2013) below.

De Soete, Lievens, & Druart (2013). Strategies for dealing with the diversity-validity dilemma in personnel selection: Where are we and where should we go? *Journal of Work and Organizational Psychology, 29,* 3-12. http://dx.doi.org/10.5093/tr2013a2

Methods determined most useful:

- Simulation-based assessments
- Alternative cognitive measures (e.g., logic-based)
- Statistical procedures

Sackett, P. R., Schmitt, N., Ellingson, J. E., Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-Affirmative-Action world. *American Psychologist, 56*(4), 302-318.

Recommendations:

- Assess all relevant attributes
    - organizational interests and performance goals
- Minimize verbal content
- Provide test preparation
- Employ face-valid assessments
- Measure experiences

Conclusion: Both diversity and performance cannot be maximized.

Outtz, J. L. (Ed.). (2010). *Adverse impact: Implications for organizational staffing and high stakes selection.* New York, NY: Taylor & Francis.

Most relevant chapters:

Goldstein, H. W., Scherbaum, C. A., & Yusko, K. P. (2010). Revisiting *g*: Intelligence, adverse impact, and personnel selection. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 95-134). New York, NY: Taylor & Francis.

Murphy, K. R. (2010). How a broader definition of the criterion domain changes our thinking about adverse impact. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 137-160). New York, NY: Taylor & Francis.

Outtz, J. L., & Newman, D. A. (2009). A theory of adverse impact. In J. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes testing* (pp. 53–93). New York, NY: Taylor & Francis.

Sackett, P. R., & Shen, W. (2010). Subgroup differences on cognitive tests in contexts other than personnel selection. In J. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 323-348). New York, NY: Taylor & Francis.

Sackett, P. R., De Corte, W., & Lievens, F. (2009). Decision aids for addressing the validity–adverse impact trade-off. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 453–472). New York, NY: Taylor & Francis.

Schmitt, N., & Quinn, A. (2009). Reductions in measured subgroup mean differences: What is possible? In J. L. Outtz (Ed.), *Implications of organizational staffing and high stakes selection* (pp. 425–451). New York, NY: Taylor & Francis Group.

REFERENCES

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice,* 25(*4*), 36-46.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14(*3*), 219–234. https://doi.org/10.1207/S15324818AME1403_2

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Ackerman, T. (1988). *An explanation of differential item functioning from a multidimensional perspective*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Alliger, G. M., & Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and Psychological Measurement, 60*, 59-72.

Alonso, P., Moscoso, S., & Salgado, J. F. (2017). Structured behavioral interview as a legal guarantee for ensuring equal employment opportunities for women: A meta-analysis. *The European Journal of Psychology Applied to Legal Context, 9,* 15-23.

Alonso, P., Moscoso, S., & Salgado, J. F. (2017). Structured behavioral interview as a legal guarantee for ensuring equal employment opportunities for women: A meta-analysis. *The European Journal of Psychology Applied to Legal Context, 9,* 15-23.

Alptekin, C. (2006). Cultural familiarity in inferential and literal comprehension in L2 reading. *System*, *34*(4), 494–508. https://doi.org/10.1016/j.system.2006.05.003

Anastasi A, Urbina S. (1997). *Psychological testing*. Upper Saddle River, NJ: Prentice Hall.

Anderson, L. M., Hoffman, R., Tate, B., Jenkins, J., Parish, C., Stachowski, A., & Dressel, J. D. (2011). *Assessment of assembling objects (AO) for improving predictive performance of the Armed Forces Qualifications Test*. (Tech. Rep. No. 1282). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior, 17,* 1-12.

Appel, M., Weber, S. & Kronberger, N. (2015). The influence of stereotype threat on immigrants: A meta-analytic review. *Frontiers in Psychology, 6,* 900. https://doi.org/10.3389/fpsyg. 2015.00900

Aronson, J., & McGlone, M. (2009). Stereotype and social identity threat. In T. Nelson (Ed.), *The handbook of prejudice, stereotyping, and discrimination.* New York: Guilford

Arth, T. O., & Skinner, J. (1986). *Aptitude selection for Air Force officer non-aircrew jobs.* Paper presented at the annual meeting of the Military Testing Association, Mystic, CT.

Arth, T.O. (1986). *Validation of the AFOQT for non-rated officers,* AFHRL-TP-85-50. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower, and Personnel Division.

Arthur, W., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology*, *55*(4), 985–1008. https://doi.org/10.1111/j.1744-6570.2002.tb00138.x

Arthur, W., Jr., Doverspike, D., Barrett, G. V., & Miguel, R. (2013). Chasing the title VII Holy Grail: The pitfalls of guaranteeing adverse impact elimination. *Journal of Business and Psychology, 28,* 473-485.

Arthur, W., Jr., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology*, 99(*3*), 535–545. https://doi.org/10.1037/a0035788

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Are Black-White differences in survey results due to response styles? *Public Opinion Quarterly, 48,* 409-427.

Bachman, J. G., O'Malley, P. M., & Freedman-Doan, P. (2010). Response styles revisited: Racial/ethnic and gender differences in extreme responding (Monitoring the Future Occasional Paper No. 72). Ann Arbor, MI: Institute for Social Research. Available from http:// www.monitoringthefuture.org/

Banks, K. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *Applied Measurement in Education, 19*(2), 115-132.

Barron, L. G., & Rose, M. R. (2013). Relative Validity of Distinct Spatial Abilities: An example with implications for diversity: Spatial Abilities: Validity and Diversity. *International Journal of Selection and Assessment*, *21*(4), 400–406. https://doi.org/10.1111/ijsa.12049

Batchelor, J. H., & Miao, C. (2016). Extreme response style: A meta-analysis. *Journal of Organizational Psychology, 16*(2), 51-62.

Beck, M. (2015, Apr 16). U.S. news: Medical-school test gets a revamp --- new version of entrance exam probes aspiring doctors on subjects beyond the core science. *Wall Street Journal.*

Benson, N., Kranzler, J. H., & Floyd, R. G. (2018). Exploratory and confirmatory factor analysis of the Universal Nonverbal Intelligence Test-Second Edition: Testing dimensionality and invariance across age, gender, race, and ethnicity. *Assessment. https://doi.org/10.1177/1073191118786584*

Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology, 41,* 174-181.

Berger, F. R., Gupta, W. B., & Berger, R. M. (1990). *Air Force officer qualifying test (AFOQT) Form P: Test manual.* Psychometrics Inc: Sherman Oaks CA.

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias. Baltimore*: Johns Hopkins University Press.

Bernardin, H. J., & Cooke, D. K. (1993). Validity of an honesty test in predicting theft among convenience store employees. *Academy of Management Journal, 36*, 1097–1108.

Berry, C. M., Sackett, P. R., & Wiemann, S. (2007). A review of recent developments in integrity test research. *Personnel Psychology, 60,* 271–301.

Bhardwaj, S., Hooda, M., & Rathee, N. K. (2018). Hypercompetitive attitude among athletes: A behavioral analysis. *European Journal of Physical Education and Sport Science, 4*(5), 49-57.

Bobko P, Roth PL, Potosky D. (1999). Derivation and implications of a meta-analysis matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 561–589.

Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on Black-White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology, 66*(1), 91–126.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 561-589.

Bosco, A., Longoni, A. M., & Vecchi, T. (2004). Gender effects in spatial orientation: Cognitive profiles and mental strategies. *Applied Cognitive Psychology, 18,* 519-532.

Bowen, N. K., Wegmann, K. M., & Webber, K. C. (2013). Enhancing a brief writing intervention to combat stereotype threat among middle-school students. *Journal of Educational Psychology*, *105*(2), 427.

Braonsford, J. D., & Johnson, M. K. (1973). Considerations of some problems of comprehension. In W. G. Chase (Ed.), *Visual information processing.* Orlando, FL: Academic Press.

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement, 41*(4), 291-310.

Brouwers, S. A., Van de Vijver, F. J. R., & Van Hemert, D. A. (2009). Variation in Raven's progressive matrices scores across time and place. *Learning and Individual Differences, 19,* 330-338.

Budoff, M., Gimon, A., & Corman, L. (1974). Learning potential measurement with Spanish-speaking youth as an alternative to IQ tests: A first report. *Interamerican Journal of Psychology, 8,* 233-246.

Cadinu, M., & Galdi, S. (2012). Gender differences in implicit gender self-categorization lead to stronger gender self-stereotyping by women than by men. *European Journal of Social Psychology, 42,* 546–551.

Cadinu, M., Latrofa, M., & Carnaghi, A. (2013). Comparing self-stereotyping with in-group-stereotyping and out-group stereotyping in unequal-status groups: The case of gender. *Self and Identity, 12*, 582-596.

Cai, H., Sedikides, C., Gaertner, L., Wang, C., Carvallo, M., Xu, Y., O'Mara, E. M., & Jackson, L. E. (2011). Tactical self-enhancement in China: Is modesty at the service of self-enhancement in East-Asian culture? *Social Psychological and Personality Science 2,* 59–64.

Camara, W. J., & Schmidt, A. E. (1999). *Group differences in standardized testing and social stratification* (College Board Report No. 99-5). New York: College Entrance Examination Board.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50,* 655–702.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50,* 655–702.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*(7), 958-975.

Capretta Raymond, C. (2006). *The new realities of identifying high potentials*. Presented at the Conference Board Succession Management Conference, New York.

Carlton, S., & Harris, A. (1992). *Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons* (No. RR-92-64). Princeton, NJ: Educational Testing Service.

Carpenter, P. A., & Just, M. A. (1986). Cognitive processes in reading. *Reading Comprehension: From Theory to Practice*, 11–29.

Carretta, T. R. (2010). Air Force Officer Qualifying Test validity for non-rated officer specialties, *Military Psychology, 22,* 450-464.

Carretta, T. R. (2013). Predictive validity of pilot selection instruments for remotely piloted aircraft training outcome. *Aviation, Space, and Environmental Medicine, 84,* 47-53.

Carretta, T. R., & Ree, M. J. (2003). Pilot selection methods. In B. H. Kantowitz (Series Ed.) & P. S. Tsang & M. A. Vidulich (Vol. Eds.). *Human factors in transportation: Principles and practices of aviation psychology* (pp. 357-396). Mahwah, NJ: Erlbaum.

Carretta, T. R., Rose, M. R., & Trent, J. D. (2016). *Air Force Officer Qualifying Test Form T: Initial Item-, Test-, Factor-, and Composite-Level Analyses*. 711 Human Performance Wing Wright-Patterson AFB United States.

Carretta. T. R. (2008). *Predictive validity of the Air Force Officer Qualifying Test for USAF air battle manager training performance,* AFRL-RH-WP-TR-2009-0007. Wright-

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge, UK: Cambridge University Press.

Casey, M. B., Nuttall, R. L., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. Developmental Psychology, 33(4), 669–680. https://doi.org/10.1037/0012-1649.33.4.669

Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental psychology*, *31*(4), 697.

Chan, D., & Shmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143-159.

Chen, Y-F., Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 Reading Assessment. *Educational Assessment, 19*(2), 77-96.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63, 83–117. doi:10.1111/j.1744-6570.2009.01163.x

Church, A. H., & Desrosiers, E. I. (2006). *Talent management: Will the high potentials please stand up.* Symposium presented at the Society for Industrial and Organizational Psychology Conference, Dallas.

Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1075 (1991).

Cohen, G. L., Garcia, J., Apfel, N., & Máster, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313,* 1307-310. 10.1126/science.1128317

Colella, A., Hebl, M., & King, E. (2017). One hundred years of discrimination research in the *Journal of Applied Psychology:* A sobering synopsis. *Journal of Applied Psychology, 102*(3), 500-513.

College Board. (2016). *2016 College-bound seniors: Total group profile report.* New York: College Entrance Examination Board.

Connolly, J. (2001). *Assessing the construct validity of a measure of learning agility* (Doctor of Philosophy Psychology, Florida International University). https://doi.org/10.25148/etd.FI14060893

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53*(2), 325-351.

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53*(2), 325-351.

Costa, P. T., & McCrae, R. R. (1992). The Five-Factor Model of Personality and Its Relevance to Personality Disorders. *Journal of Personality Disorders*, *6*(4), 343–359. https://doi.org/10.1521/pedi.1992.6.4.343

Costa, P. T., Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*(2), 322–331. https://doi.org/10.1037/0022-3514.81.2.322

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York, NY: HarperCollins.

Danaher, K., & Crandall, C. S. (2008). Stereotype Threat in Applied Settings Re-Examined. *Journal of Applied Social Psychology*, *38*(6), 1639–1655. https://doi.org/10.1111/j.1559-1816.2008.00362.x

Davis, D., Dorsey, J. K., Franks, R. D., Sackett, P. R., Searcy, C. A., & Zhao, X. (2013). Do racial and ethnic group differences in performance on the MCAT exam reflect test bias? *Academic Medicine, 88(*5), 593-602.

De Meijer, L. A. L., Born, M. Ph., Terlouw, G., & Molen, H. T. van der. (2006). Applicant and Method Factors Related to Ethnic Score Differences in Personnel Selection: A Study at the Dutch Police. *Human Performance*, *19*(3), 219–251. https://doi.org/10.1207/s15327043hup1903_3

De Meuse, K. P., Dai, G., Eichinger, R. W., Page, R. C., Clark, L. P., & Zewdie, S. (2011). The development and validation of a self-assessment of learning agility. *Society for Industrial and Organizational Psychology Conference, Chicago, Illinois*.

De Soete, B., Lievens, F., & Druart, C. (2012). An update on the diversity-validity dilemma in personnel selection: A review. *Psychological Topics, 21*(3)*,* 399-424.

De Soete, B., Lievens, F., & Druart, C. (2013). Strategies for dealing with the diversity-validity dilemma in personnel selection: Where are we and where should we go? *Revista de Psicología Del Trabajo y de Las Organizaciones*, *29*(1), 3–12. https://doi.org/10.5093/tr2013a2

De Soete, B., Lievens, F., Oostrom, J., & Westerveld, L. (2013). Alternative Predictors for Dealing with the Diversity-Validity Dilemma in Personnel Selection: The constructed response multimedia test: Alternative Predictors in Personnel Selection. *International Journal of Selection and Assessment*, *21*(3), 239–250. https://doi.org/10.1111/ijsa.12034

Dean, M. A. (2013). Examination of ethnic group differential responding on a biodata instrument*. Journal of Applied Social Psychology, 43,* 1905–1917.

Dollinger, S. J., & Clark, M. H. (2012). Test-taking strategy as a mediator between race and academic performance*. Learning and Individual Differences,* 22(*4*), 511–517. https://doi.org/10.1016/j.lindif.2012.03.010

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied psychological measurement*, *28*(4), 227-246.

Dorans, N. J., & Zeller, K. (2004). Examining Freedle's Claims About Bias and His Proposed Solution: Dated Data, Inappropriate Measurement, and Incorrect and Unfair Scoring. *ETS Research Report Series*, *2004*(2), 1-33.

Dossey, J. A. (1993). *Can Students Do Mathematical Problem Solving? Results from Constructed-Response Questions in NAEP's 1992 Mathematics Assessment*. US Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328.

Doyle, R. A., & Voyer, D. (2016). Stereotype manipulation effects on math and spatial test performance: A meta-analysis. *Learning and Individual Differences*, *47*, 103–116. https://doi.org/10.1016/j.lindif.2015.12.018

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

Educational Testing Service. (2018). *A Snapshot of the individuals who took the GRE General Test*. Princeton, NJ: Educational Testing Service.

Edwards, B. D., & Arthur, W. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology*, *92*(3), 794–801. https://doi.org/10.1037/0021-9010.92.3.794

Ellis, A. P., & Ryan, A. M. (2003). Race and Cognitive-Ability Test Performance: The Mediating Effects of Test Preparation, Test-Taking Strategy Use and Self-Efficacy. *Journal of Applied Social Psychology*, 33(*12*), 2607-2629.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136,* 103-127.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). *Uniform guidelines on employee selection procedures.* Federal Register, 43, 38290–39315.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*(2), 24-35.

Fagan, J. F., & Holland, C. R. (2002). Equal opportunity and racial differences in IQ. *Intelligence*, *30*(4), 361-387.

Fagan, J., & Holland, C. R. (2007). Racial equality in intelligence: Predictions form a theory of intelligence as processing. *Intelligence, 35*(4), 319-334.

Farh, C. I. C. C., Seo, M.-G., & Tesluk, P. E. (2012). Emotional intelligence, teamwork effectiveness, and job performance: The moderating role of job context. *Journal of Applied Psychology, 97*(4), 890–900. https://doi.org/10.1037/a0027377

Feng, J., Spence, I. & Pratt, J. (2007). Playing an action game reduces gender differences in spatial cognition. *Psychological Science, 18*, 850–855.

Ferreter, J., Goldstein, H., Scherbaum, C., Yusko, K., & Jun, H. (2008, April). *Reducing adverse impact using a nontraditional cognitive ability assessment.* Poster presented at the Society for Industrial and Organizational Psychology 23rd Annual Conference, San Francisco.

Finegold, L., & Rogers, D. (1985). *Relationship between Air Force Officer Qualifying Test scores and success in air weapons controller training,* AFHRL-TR-85-13. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five U.S. racial groups. *Personnel Psychology, 61(3),* 579–616. https://doi.org/10.1111/j.1744-6570.2008.00123.x

Forscher, P. S., Taylor, V. J., Cavagnaro, D., Lewis, N. A., Moshontz, H., Mark, A. Y., Appleby, S., Batres, C., Bennett-Day, B., Buchanan, E. M., Chopik, W. J., Damian, R. I., Ellis, C. E., Faas, C., Gaither, S., Green, D., Hall, B. F., Hinojosa, B. M., Howell, J. L., … Chartier, C. R. (2019). A Multi-Site Examination of Stereotype Threat in Black College Students Across Varying Operationalizations [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/6hju9

Freedle, R. O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review, 73,* 1-42.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment, 6,* 115–123. http://dx.doi.org/10.1111/ 1468-2389.00080

Gadermann, A. M., Chen, M. Y., Emerson, S. D., & Zumbo, B. D. (2018). Examining Validity Evidence of Self-Report Measures Using Differential Item Functioning: An Illustration of Three Methods. *Methodology*, *14*(4), 165–176. https://doi.org/10.1027/1614-2241/a000156

Gallagher, A. M. (1992). *Sex differences in problem-solving used by high-scoring examinees on the SAT-M (*College Board Report No. 92-2, ETS RR No. 92-33). New York: College Board Publications.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75*(3), 165-190. https://doi.org/10.1006/jecp.1999.2532

Gallagher, A., Levin, J., & Cahalan, C. (2002). *Cognitive Patterns of Gender Differences on Mathematics Admissions Tests.* GRE Board Professional Report No. 96-17, Princeton, NJ: Educational Testing Service.

Gallagher, A., Levin, J., & Cahalan, C. (2002). *GRE research: Cognitive patterns of gender differences on mathematics admissions tests* (ETS Report No. 0219). Princeton, NJ: Educational Testing Service.

Gandy, J. A., Dye, D. A., & MacLane, C. N. (1994). Federal government selection: The individual achievement record. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (p. 275–309). CPP Books.

Ganley, C. M., Vasilyeva, M., & Dulaney, A. (2014). Spatial ability mediates the gender difference in middle school students' science performance. *Child Development, 85*(4), 1419-1432.

Geary, D. C., Salthouse, T. A., Chen, G.-P., & Fan, L. (1996). Are East Asian versus American differences in arithmetical ability a recent phenomenon? *Developmental Psychology, 32*(2), 254-262.

Geisinger, K. F. (1988). The golden rule in psychological testing: Please, please don't do it unto me. *Theoretical & Philosophical Psychology, 8*(2), 15–23. https://doi.org/10.1037/h0091444

Gillespie, J. Z., Converse, P. D., & Kriska, S. D. (2010). Applying Recommendations from the Literature on Stereotype Threat: Two Field Studies. *Journal of Business and Psychology*, *25*(3), 493–504. https://doi.org/10.1007/s10869-010-9178-1

Glück, J., & Fitting, S. (2003). Spatial strategy selection: Interesting incremental information. *International Journal of Testing, 3*(3), 293-308.

Goldstein, H. W., Yusko, K. P., & Nicolopoulos, V. (2001). Exploring black-white subgroup differences of managerial competencies. *Personnel Psychology, 54,* 783–807.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential Item Functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104-109.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*(6), 645–662. https://doi.org/10.1016/j.appdev.2003.09.002

Guay, R. B. (1980). *Spatial ability measurement: a critique and an alternative.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED189166).

Guimond, S., Chatard, A., Martinot, D., Crisp, R. J., & Redersdorff, S. (2006). Social comparison, self-stereotyping, and gender differences in self-construals. *Journal of Personality and Social Psychology, 90,* 221–242.

Hanges, P. J., & Feinberg, E. G. (2010). International perspectives on adverse impact: Europe and beyond. In J. L. Outtz (Ed.), *Adverse Impact: Implications for Organizational Staffing and High Stakes Selection* (pp. 349-373). New York, NY: Taylor & Francis Group.

Harold, C. M., McFarland, L. A., & Weekley, J. A. (2006). The validity of verifiable and nonverifiable biodata items: An examination across applicants and incumbents. *International Journal of Selection and Assessment, 14,* 336–346.

Hartke, D. D., & Short. L. O. (1988). Validity of the academic aptitude composite of the Air Force Officer Qualifying Test (AFOQT), AFHRL-TP-87-61. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower, and Personnel Division.

Harzing, A. W. K., Brown, M., Köster, K., & Zhao, S. (2012). Response style differences in cross-national research. *Management International Review, 52,* 341-363.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Gerrard, M. O. M. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92,* 373-385.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high scoring individuals. *Science,* 269, 41–45.

Hedlund, J., Wilt, J. M., Nebel, K. L., Ashford, S. J., & Sternberg, R. J. (2006). Assessing practical intelligence in business school admissions: A supplement to the graduate management admissions test. *Learning and Individual Differences, 1*6(2), 101–127. https://doi.org/10.1016/j.lindif.2005.07.005

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence, 32,* 175-191.

Held, J. D., & Carretta, T. R. (2013). *Evaluation of tests of processing speed, spatial ability, and working memory for use in military occupational classification*. (NPRST-TR-14-1). Millington, TN: Navy Personnel Research, Studies, and Technology.

Hessels, M. G. P., & Hamers, J. H. M. (1993). A Learning potential test for ethnic minorities. In J. H. M. Hamers, K. Sijtsma & A. J. J. M. Ruijssenaars (Eds.), Learning potential assessment: Theoretical, methodological, and practical issues (pp. 285-311). Lisse, Netherlands: Swets & Zeitlinger.

Hirnstein, M., Bayer, U., & Hausmann, M. (2009). Sex-specific response strategies in mental rotation. *Learning and Individual Differences*, *19*(2), 225–228. https://doi.org/10.1016/j.lindif.2008.11.006

Hough, L. M., & Ones, D. S. (2002). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In A. Neil and O.S. Ones (Eds.), *International Handbook of Work and Organizational Psychology* (pp. 233-277). Thousand Oaks, CA: Sage.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, Detection and Amelioration of Adverse Impact in Personnel Selection Procedures: Issues, Evidence and Lessons Learned. *International Journal of Selection and Assessment*, *9*(1 & 2), 152–194. https://doi.org/10.1111/1468-2389.00171

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83,* 179–189.

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83,* 179–189.

Huffcutt, A. I., & Roth, P. O. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83,* 179–189.

Hugdahl, K., Thomsen, T., & Ersland, L. (2006). Sex differences in visuo-spatial processing: an fMRI study of mental rotation. *Neuropsychologia*, *44*(9), 1575-1583.

Huguet, P., & Regner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of Educational Psychology, 99*(3), 545.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*(1), 53-69.

Hyde, J. S., & Linn, M. C. (2006). Gender similarities in mathematics and science. *Science, 314,* 599-600.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321,* 494-495.

Imus, A., Schmitt, N., Kim, B., Oswald, F. L., Merritt, S., & Wrestring, A. F. (2011). Differential item functioning in biodata: Opportunity access as an explanation of gender- and race-related DIF. *Applied Measurement in Education, 24*(1), 71-94. http://dx.doi.org/10.1080/08957347.2011.532412

Ison, D.C., Herron, R., & Weiland, L. (2016). Two decades of progress for minorities in aviation. *Journal of Aviation Technology and Engineering, 6*(1), 25-33.

Janssen, A. B., & Geiser, C. (2010). On the relationship between solution strategies in two mental rotation tasks. *Learning and Individual Differences, 20*(5), 473-478.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Johnson, T., & F., Van de Vijver. (2003). Social desirability bias in cross cultural research. In J. Harkness, F. Van de Vijver, & P. Mohler Hoboken (Eds.), *cross-cultural survey methods* (pp. 195-204). Hoboken, NJ: John Wiley & Sons.

Johnson, T., Kulesa, P., Cho, Y., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36,* 264–277.

Jordan, K., Wüstenberg, T., Heinze, H. J., Peters, M., & Jäncke, L. (2002). Women and men exhibit different cortical activation patterns during mental rotation tasks. *Neuropsychologia*, *40*(13), 2397-2408.

Karami, H., & Salmani Nodoushan, M. A. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies, 5*(3), 133-142.

Kato, K., Moen, R., & Thurlow, M. (2009). Differentials of a state reading assessment: Item functioning, distractor functioning, and omission frequency for disability categories. *Educational Measurement: Issues and Practice*, *28*(2), 28–40

Kaufman, S. B., & Sternberg, R. J. (2007). *Giftedness in Euro-American culture*. In S. N. Phillipson & M. McCann (Eds.), Conceptions of giftedness: Socio-cultural perspectives. Mahwah, NJ: Lawrence Erlbaum

Keiser, H. N., Sackett, P. R., Kuncel, N. R., & Brothen, T. (2016). Why women perform better in college than admission scores would predict: Exploring the roles of conscientiousness and course-taking patterns. *Journal of Applied Psychology, 101*(4), 569–581. https://doi.org/10.1037/apl0000069

Keller, G. D., Deneen, J. R., & Magallán, R. J. (Eds.). (1991). *Assessment and access: Hispanics in higher education*. SUNY Press.

Kendall, I. M., Verster, M. A., & Von Mollendorf, J. W. (1988). *Test performance of Blacks in Southern Africa.* In S. H. Irvine & J. W. Berry (Eds.), Human abilities in cultural context (p. 299–339). Cambridge University Press. https://doi.org/10.1017/CBO9780511574603.013

Keshavarz, M. H., Atai, M. R., & Ahmadi, H. (2007). Content schemata, linguistic simplification, and EFL readers' comprehension and recall. *Reading in a Foreign Language, 19,* 19-33.

Kevelson, M. J. (2019). *The Measure Matters: Examining Achievement Gaps on Cognitively Demanding Reading and Mathematics Assessments* (Research Report Series No. RR-19-43). Princeton, NJ: Educational Testing Service.

Kiddler, W. C., & Rosner, J. (2002). How the SAT creates built-in-headwinds: An educational and legal analysis of disparate impact. *Santa Clara Law Review, 43,* 131-212.

Kluger, A. N., & Colella, A. (1993). Beyond the mean bias: The effect of warning against faking on biodata item variances. *Personnel Psychology, 46,* 763–780.

Kobrin, J. L., Sathy, V., Shaw, E., J., (2007). *A historical view of subgroup performance differences on the SAT Reasoning Test*. New York: College Board.

Kock, F. D., & Schlechter, A. (2009). Fluid intelligence and spatial reasoning as predictors of pilot training performance in the South African Air Force (SAAF). *SA Journal of Industrial Psychology*, *35*(1), 31-38.

Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition, 29*(5), 745–756.

Kurman, J. (2003). Why is self-enhancement low in certain collectivist cultures? An investigation of two competing explanations. *Journal of Cross-Cultural Psychology, 34,* 496-510.

Kurman, J., & Sriram, N. (2002). Interrelationships among vertical and horizontal collectivism, modesty, and self-enhancement. *Journal of Cross-Cultural Psychology, 33*(5), 71-86.

Lalwani, A. K., Shavitt, S., & Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding? *Journal of Personality and Social Psychology, 90,* 165–178.

Lalwani, A. K., Shrum, L. J., & Chiu, C. (2009). Motivated response styles: The role of cultural values, regulatory focus, and self-consciousness in socially desirable responding. *Journal of Personality and Social Psychology, 96,* 870–882.

Larson, E. C. (2019). A Meta-Analysis of Information Processing Measures of Intelligence, Performance, and Group Score Differences.

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology, 67,* 241-293.

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology, 67,* 241-293.

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology, 67,* 241-293.

Lewis & Lewis, N. A., & Michalak, N. M. (2019). *Has Stereotype Threat Dissipated Over Time? A Cross-Temporal Meta-Analysis* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/w4ta2

Lewis, J. D., DeCamp-Fritson, S. S., Ramage, J. C., McFarland, M. A., & Archwamety, T. (2007). Selecting for Ethnically Diverse Children Who May Be Gifted Using Raven's Standard Progressive Matrices and Naglieri Nonverbal Abilities Test. *Multicultural Education*, *15*(1), 38-42.

Lewis, N. A., & Michalak, N. M. (2019). *Has Stereotype Threat Dissipated Over Time? A Cross-Temporal Meta-Analysis* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/w4ta2

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. Psicologica, 30, 343–370.

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

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

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology*, 97(*2*), 460–468. https://doi.org/10.1037/a0025741

Lievens, F., De Corte, W., & Westerveld, L. (2015). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management, 41*(6), 1604-1627.

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., & De Soete, B. (2019). Constructed response formats and their effects on minority–majority differences and validity. *Journal of Applied Psychology*, 104(*5*), 715–726. https://doi.org/10.1037/apl0000367

Linn, R. L., & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational Measurement: Issues and Practice, 6*(2), 13–17. https://doi.org/10.1111/j.1745-3992.1987.tb00405.x

Loewen, J. W., Rosser, P., & Katzman, J. (1988). *Gender bias in SAT items.* New Orleans, LO: American Educational Research Association.

Lohman, D. F., & Gambrell, J. L. (2012). Using nonverbal tests to help identify academically talented children. *Journal of Psychoeducational Assessment*, *30*(1), 25–44.

Lohman, D. F., Korb, K., & Lakin, J. (2008). Identifying academically gifted English language learners using nonverbal tests: A comparison of the Raven, NNAT, and CogAT. Gifted Child Quarterly, 52, 275-296. (Research Paper of the Year Award from the National Association of Gifted Children)

Lombardo, M. M., & Eichinger, R. W. (2000). High potentials as high learners. *Human Resource Management*, *39*(4), 321–329.

Loreozi-Cioldi, F. (1991). Self-stereotyping and self-enhancement in gender groups. *European Journal of Social Psychology, 21*, 403-417.

Maeda, Y., & Yoon, S. Y. (2013). A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: Visualization of rotations (PSVT:R). *Educational Psychology Review, 25,* 69-94.

Malda, M., van de Vijver, F. J., & Temane, Q. M. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence*, *38*(6), 582-595.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). A review of recent developments in differential item functioning. *ETS Research Report Series*, *2008*(2), 1-32.

Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. Personality and Social Psychology Bulletin, 28(9), 1183-1193.

Mattern, K., Sanchez, E., & Ndum, E. (2017). Why do achievement measures underpredict female academic performance?. *Educational Measurement: Issues and Practice*, *36*(1), 47-57.

McDaniel, M. A., & Weekley, J. A. (2012, April). *Controlling for elevation and scatter in situational judgment test scoring: A replication.* Paper presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

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

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, 96(*2*), 327–336. https://doi.org/10.1037/a0021983

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812–821.

McGee, M. G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, *86*(5), 889.

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), Educational measurement (pp. 13-103). New York: MacMillan.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied psychological measurement, 17(4), 297-334.

Moreau, D., Clerc, J., Mansy-Dannay, A., & Guerrien, A. (2012). Enhancing spatial ability through sport practice: Evidence for an effect of motor training on mental rotation performance. *Journal of Individual Differences, 33*(2), 83-88.

Mount, M. K., Witt, L. A, & Barrick, M. R. (2000). Incremental validity of empirically keyed biodata scales over GMA and the five factor personality constructs. *Personnel Psychology, 53,* 299–323.

Nadler, J. T., & Clark, M. H. (2011). Stereotype Threat: A Meta-Analysis Comparing African Americans to Hispanic Americans 1. *Journal of Applied Social Psychology*, *41*(4), 872-890.

Naglieri, J. A. (2005). The cognitive assessment system. In D. P. Flanagan and P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 441-460). New York, NY: Guilford Press.

Naglieri, J. A., & Ford, D. Y. (2003). Addressing underrepresentation of gifted minority children using the Naglieri Nonverbal Ability Test (NNAT*). Gifted Child Quarterly, 47*, 155-160.

Naglieri, J. A., & Ronning, M. E. (2000). Comparison of White, African American, Hispanic, and Asian children on the Naglieri Nonverbal Ability Test. *Psychological assessment*, *12*(3), 328.

National Assessment of Educational Progress (NAEP), 2015. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

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

Nelson, L. C. (2003). *Working memory, general intelligence, and job performance* [Unpublished dissertation]. University of Minnesota.

Newman, D. A., & Lyon, J. S., (2009). Recruitment efforts to reduce adverse impact: Targeted recruiting for personality, cognitive ability, and diversity. *Journal of Applied Psychology, 94*(2), 298-317.

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314–1334. https://doi.org/10.1037/a0012702

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching, 49*(6), 778-803.

Nordvik, H., & Amponsah, B. (1998). Gender differences in spatial abilities and spatial activity among university students in an egalitarian educational system. *Sex Roles, 38,* 1009-1023.

O'Neill, K., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillside, NJ: Erlbaum.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79*, 845-849.

Olenick, J., Bhatia, S. and Ryan, A.M. (2016), Effects of g-loading and time lag on retesting in job selection. *International Journal of Selection and Assessment, 24*, 324-336. https://doi.org/10.1111/ijsa.12151

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11,* 245-271.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679–703.

Outtz, J. L. (Ed.). (2010). *Adverse impact: Implications for organizational staffing and high stakes selection.* New York, NY: Taylor & Francis.

Outtz, J. L., & Hanges, P. J. (2013). *Barrier Analysis of the Air Traffic Control Specialists (ATCS) Centralized Hiring Process.* Washington, DC: Outtz and Associates. Prepared for the Federal Aviation Administration. Retrieved from https://www.faa.gov/

Pangallo, A., Zibarras, L., & Patterson, F. (2016). Measuring resilience in palliative care workers using the situational judgement test methodology. *Medical Education, 50,* 1131-1142.

Patterson AFB, OH: Air Force Research Laboratory, Human Effectiveness Directorate, Crew Systems Interface Division, Supervisory Control Interfaces Branch.

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, *45*(3), 247–269

Picho, K., Rodriguez, A., & Finnie, L. (2013). Exploring the moderating role of context on the mathematics performance of females under stereotype threat: A Meta-Analysis. *The Journal of Social Psychology*, *153*(3), 299–333. https://doi.org/10.1080/00224545.2012.737380

Pierce, L. G., Broach, D., Byrne, C. L., & Bleckley, M. K. (October 2014). *Using biodata to select Air Traffic Controllers.* (DOT/FAA/AM-14/8). Oklahoma City, OK: FAA Civil Aerospace Medical Institute.

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

Prifitera, A., Saklofske, D. H., Weiss, L. G., & Rolfhus, E. (2005). The WISC-IV in the Clinical Assessment Context. In A. Prifitera, D. H. Saklofske, L. G. Weiss & E. Rolfhus (Eds.), WISC-4 Clinical use and interpretation (pp. 3-32). San Diego, CA: Elsevier Academic Press.

Pyburn, K. M., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology, 61,* 143-151.

Randall, J. G. (2012). *Is retest bias biased? An examination of race, sex, and ability differences in retest performance on the Wonderlic personnel test* (Master's thesis). https://scholarship.rice.edu/handle/1911/8299

Randall, J. G., Villado, A. J., & Zimmer, C. U. (2016). Is retest bias biased? Examining race and sex differences in retest performance. *Journal of Personnel Psychology, 15*(2), 45-54.

Randall, J.G., & Villado, A. J. (2017). Take two: Sources and deterrents of score change in employment testing. *Human Resource Management, 27,* 536–553.

Raven, J. C. (1938). *Progressive Matrices: A perceptual test of intelligence, 1938, sets A, B, C, D, and E.* London: H. K. Lewis.

Reeve, C. L., & Lam, H. (2007). The relation between practice effects, test-taker characteristics, and degree of *g*-saturation. *International Journal of Testing, 7*(2), 225-242.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35,* 1-62.

Resing, W. C. M., Tunteler, E., de Jong, F. M., Bosma, T. (2009). Dynamic testing in indigenous and ethnic minority children. *Learning and Individual Differences, 19*(4), 445-450. https://doi.org/10.1016/j.lindif.2009.03.006

Ridgway, T. (1997). Thresholds of background knowledge effect in foreign language reading. *Reading in a Foreign Language, 11,* 151–166.

Roberts, H. E., & Skinner, J. (1996). Gender and racial equity of the Air Force Officer Qualifying Test in officer training school selection decisions. Military Psychology, 8, 95- 113

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. Journal of Educational Measurement, 40, 163–184. http://dx.doi.org/10 .1111/j.1745-3984.2003.tb01102.x

Rosenthal, H. E. S., & Crisp, R. J. (2006). Reducing stereotype threat by blurring intergroup boundaries. *Journal of Personality and Social Psychology, 32,* 501–511

Roth P, Bobko P, McFarland L, Buster M. (2008). Work sample tests in personnel selection: A meta-analysis of black-white differences in overall and exercise scores. PERSONNEL PSYCHOLOGY, 61, 637–661.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297-330.

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., & Bobko, P. (2002). Corrections for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology, 87,* 369–376.

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., & Bobko, P. (2002). Corrections for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology, 87,* 369–376.

Roussos, L. A., & Stout, W. F. (1996). A multidimensionality based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 353-371.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*(2), 248-269.

Russell, T. L., & Peterson, N. G. (2001). The experimental battery: Basic attribute scores for predicting performance in a population of jobs. In *Exploring the limits in personnel selection and classification* (pp. 269-306). J.P Campbell and D. J. Knapp (Eds.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ryan, A. M., & Tippins, N. T. (2004). Attracting and selecting: What psychological research tells us. *Human Resource Management, 43,* 305–318.

Sackett, P. R., & Shen, W. (2010). Subgroup differences on cognitive tests in contexts other than personnel selection. In J. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection*: 323-348. New York: Routledge.

Sackett, P. R., Eitelberg, M. J., & Sellman, W. S. (2009). *Profiles of American youth: Generational changes in cognitive skill.* Alexandria, VA: Human Resource Research Organization.

Sackett, P. R., Schmitt, N., Ellingson, J., E., & Kabin, M. B., (2001). High stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56*, 302–318.

Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. Harvard Educational Review, 80(1), 106–133.

Santelices, M. V., & Wilson, M. (2015). The revised SAT score and its potential benefits for the admission of minority students to higher education. E*ducation Policy Analysis Archives*, *23*, 113.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations (5th ed*.). San Diego, CA: Author.

Scarr, S. (1994). Culture-fair and culture-free tests. In R. J. Sternberg (Ed.), Encyclopedia of human intelligence (pp. 322-328). New York, NY: Macmillan.

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence, 67*, 44-66.

Scherbaum, C. A., & Goldstein, H. W. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement*, 68(4), 537–553. https://doi.org/10.1177/0013164407310129

Schleicher, D. J., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2010). If at first you don't succeed, try, try again: Understanding race, age, and gender differences in retesting score improvement. *Journal of Applied Psychology, 95*(4), 603–617.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology. Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmitt, A., & Dorans, N. (1988). *Differential item functioning for minority examinees on the SAT* (No. RR-88-32). Princeton, NJ: Educational Testing Service.

Schmitt, D. P., Realo, A., Voracek, M., & Jüri, A. (2008). Why can't a man be more like a woman? Sex differences in Big Five Personality Traits across 55 cultures. *Journal of Personality and Social Psychology, 94*(1), 168-182.

Schmitt, N., & Kunce, C. (2002). The effects of required elaboration of answers to biodata questions. *Personnel Psychology, 55,* 569–587.

Schmitt, N., Clause, C., & Pulakos, E. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In Cooper CL, Robertson IT (Eds.), *International Review of Industrial and Organizational Psychology, 11*, 115-139.

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., & Ramsay, L. J. (2003). Impact of elaboration on socially desirable responding and the validity of biodata measures. *Journal of Applied Psychology, 88,* 979–988.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science, 237,* 1317-1323.

Shewach, O. R., Sackett, P. R., & Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis. *Journal of Applied Psychology*, *104*(12), 1514–1534. https://doi.org/10.1037/apl0000420

Shore, C. W. (2014, October). *Reducing adverse impact for the Air Force Officer Qualifying Test: An informal report.* San Antonio, TX: Operational Technologies.

Shore, C. W., Peña, D. A., Gonzalez, M., Haight, N. R., & Wolliston, D. J. (2019). *Officer and enlisted needs analysis,* Task Order #47QFAA18F0043. San Antonio, TX: Operational Technologies.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. Journal of Personality and Social Psychology, 75(5), 1350–1362. https://doi.org/10.1037/0022-3514.75.5.1350

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (p. 45–60). Lawrence Erlbaum Associates, Inc.

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype Threat. *Annual Review of Psychology*, *67*(1), 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., Jarvin, L., & Sharpes, K. (2009). Using the theory of successful intelligence as a framework for developing assessments in AP physics. *Contemporary Educational Psychology*, *34*(3), 195-209.

Sternberg, R. J. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence*, *34*(4), 321–350. https://doi.org/10.1016/j.intell.2006.01.002

Sternberg, R.J. Sternberg Triarchic Abilities Test (Modified), Level H, 1993

Stieff, M., Dixon, B. L., Ryu, M., & Kumi, B. C. (2013). Strategy training eliminates sex differences in spatial problem solving in a STEM domain. *Journal of Educational Psychology, 106*(2), 390-402.

Stieff, M., Dixon, B. L., Ryu, M., Kumi, B. C., & Hegarty, M. (2014). Strategy training eliminates sex differences in spatial problem solving in a stem domain. *Journal of Educational Psychology*, *106*(2), 390.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology, 16*(1), 93-102.

Stokes, G. S., & Cooper, L. A. (2001). Content/construct approaches in life history form development for selection. *International Journal of Selection and Assessment, 9,* 138-151.

Stokes, G. S., Hogan, J. B., & Snell, A. F. (1993). Comparability of incumbent and applicant samples for the development of biodata keys: The influence of social desirability. *Personnel Psychology, 46,* 739–762.

Stricker, L. J., Rock, D. A., Burton, N. W., Muraki, E., & Jirele, T. J. (1994). Adjusting college grade point average criteria for variations in grading standards: A comparison of methods. Journal of Applied Psychology, 79(2), 178–183. https://doi.org/10.1037/0021-9010.79.2.178

Teachout, M., Shore, C. W., Martinez, L., & Wolliston, D. (2019). *Identifying potential measures for improved selection and classification,* Task Order Task Order #47QFAA18F0043. San Antonio, TX: Operational Technologies.

Thornton, B., Ryckman, R. M., & Gold, J. A. (2011). Hypercompetitiveness and relationships: Further implications for romantic, family, and peer relationships. *Psychology, 2*(4), 269-274.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology, 87*(6), 1020.

Tsaousis, I., Sideridis, G., & Al-Saawi, F. (2018). Differential Distractor Functioning as a Method for Explaining DIF: The Case of a National Admissions Test in Saudi Arabia. *International Journal of Testing*, *18*(1), 1-26.

United States Air Force. (2015). *Officer Qualifying Test (AFOQT) information pamphlet* [Brochure]. Author.

Uskul, A. K., Oyserman, D., & Schwarz, N. (2010). Cultural emphasis on honor, modesty, or self-enhancement: Implications for the survey response process. In M. Braun, B. Edwards, J. Harkness, T. Johnson, L. Lyberg, P. Mohler, B.E. Pennell, & T.W. Smith (Eds.), *Survey methods in multinational, multiregional and multicultural context*. Hoboken, NJ: John Wiley and Sons.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352-402.

Van Hemert, D. A., van de Vijver, F. J. R., Poortinga, Y. H., & Georgas, J. (2002). Structural and functional equivalence of the Eysenck Personality Questionnaire within and between countries. *Personality and Individual Differences, 33,* 1229-1249.

Van Iddekinge, C. H., Morgeson, F. P., Schleicher, D. J., & Campion, M. A. (2011). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology, 96*(5), 941-955.

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. Journal of Applied Psychology.

Van Iddekinge, C. H., Taylor, M. A., & Eidson, C. E. J. (2005). Broad versus narrow facets of integrity: Predictive validity and subgroup differences. *Human Performance, 18*(2), 151-177.

Villado, A. J., Randall, J. G., & Zimmer, C. U. (2016). The effect of method characteristics on retest score gains and criterion-related validity. *Journal of Business Psychology, 31,* 233-248.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*(2), 250-270.

Wainer, H., & Skorupski, W. P. (2005). Was it ethnic and social-class bias or statistical artifact? Logical and empirical evidence against Freedle's method for reestimating SAT scores. Chance, 18(2), 17–24.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*(2), 147-163.

Walker, C. M., & Gocer Sahin, S. (2017). Using a multidimensional IRT framework to better understand differential item functioning (DIF): A tale of three DIF detection procedures. *Educational and Psychological Measurement, 77*(6), 945-970.

Walters, L. C., Miller, M. R., & Ree, M. J. (1993). Structured interviews for pilot selection: No incremental validity. *The International Journal of Aviation Psychology, 3,* 25-38.

Walters, L. C., Miller, M. R., & Ree, M. J. (1993). Structured interviews for pilot selection: No incremental validity. *The International Journal of Aviation Psychology, 3,* 25-38.

Walters, L. C., Miller, M. R., & Ree, M. J. (1993). Structured interviews for pilot selection: No incremental validity. *The International Journal of Aviation Psychology, 3,* 25-38.

Walton, G. M., & Cohen, G. L. (2003). Stereotype Lift. *Journal of Experimental Social Psychology*, *39*(5), 456–467. https://doi.org/10.1016/S0022-1031(03)00019-2

Walton, G. M., & Spencer, S. J. (2009). Latent Ability: Grades and Test Scores Systematically Underestimate the Intellectual Ability of Negatively Stereotyped Students. *Psychological Science*, *20*(9), 1132–1139. https://doi.org/10.1111/j.1467-9280.2009.02417.x

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior,* 2(*1*), 295-322.

Weiss, E., Siedentopf, C., Hofer, A., Deisenhammer, E., Hoptman, M., Kremser, C., Golaszewski, S., Felber, S., Fleischhacker, W., & Delazer, M. (2003). Sex differences in brain activation pattern during a visuospatial cognitive task: A functional magnetic resonance imaging study in healthy volunteers. *Neuroscience Letters*, *344*(3), 169–172.

Weiss, L. G., Saklofske, D. H., Prifitera, A., & Holdnack, J. A. (Eds.). (2006). Wechsler Intelligence Scale for Children-4 advanced clinical interpretation. Burlington, MA: Academic Press

Wexley, K. N., Sanders, R. E., & Yukl, G. A. (1973). Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology, 57,* 233–236.

Wexley, K. N., Sanders, R. E., & Yukl, G. A. (1973). Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology, 57,* 233–236.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational Arthur Jr, W., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology, 55*(4), 985-1008.

Whitney, D. J., & Schmitt, N. (1997). Relationship between culture and responses to biodata employment items. *Journal of Applied Psychology, 82*(1), 113.

Wiley, D. E. (1990). Test validity and invalidity reconsidered. In R. Snow & D. E. Wiley (Eds.), Improving inquiry in social science (pp. 75-107). Hillsdale, NJ: Lawrence Erlbaum.

Wilson, L. D., & Zhang, L. (1998). A cognitive analysis of gender differences on constructed-response and multiple-choice assessments in mathematics test performance: A meta-analysis. *Human Performance*, 21(*3*), 291-309.

Woehr, D. J., & Huffcutt, A. I., (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189–205.

Woehr, D. J., & Huffcutt, A. I., (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189–205.

Wolfe, J. H. (1997). Incremental validity of ECAT battery factors. *Military Psychology*, *9*(1), 49-76.

Yuet, C., & Chan, H. (2003). Cultural content and reading proficiency: A comparison of Mainland Chinese and Hong Kong learners of English. *Language, Culture, and Curriculum, 16,* 60-69. https://doi.org/10.1080/07908310308666657

Yusko, K. P., & Goldstein, H. W. (2008). Siena Reasoning Test. Princeton, NJ: Siena Consulting.

Zigerell, L. J. (2017). Potential publication bias in the stereotype threat literature: Comment on Nguyen and Ryan (2008). *Journal of Applied Psychology, 102*(8), 1159–1168.