

Technical Report 1437

Exploring Personality and Performance Criterion Relationships Using the Tailored Adaptive Personality Assessment System

Editors

Deirdre J. Knapp

Human Resources Research Organization

Cristina D. Kirkendall

U.S. Army Research Institute



September 2023

**United States Army Research Institute
for the Behavioral and Social Sciences**

Approved for public release; distribution is unlimited.

**U.S. Army Research Institute
for the Behavioral and Social Sciences**

**Department of the Army
Deputy Chief of Staff, G1**

Authorized and approved:

**SCOTT B. SHADRICK, Ph.D.
Senior Scientific Technical Manager**

Research accomplished under contract
for the Department of the Army by:

Human Resources Research Organization

Technical Review by:

Melissa Glorioso, U.S. Army Research Institute
Karly Schleicher, U.S. Army Research Institute

DISPOSITION

This Technical Report has been submitted to the
Defense Technical Information Center (DTIC).

REPORT DOCUMENTATION PAGE				
1. REPORT DATE (Month Year) September 2023		2. REPORT TYPE Interim		3. DATES COVERED (Month Year)
				START DATE April 2019
				END DATE March 2021
4. TITLE AND SUBTITLE Exploring Personality and Performance Criterion Relationships Using the Tailored Adaptive Personality Assessment System				
5a. CONTRACT NUMBER W911NF-19-C-0032		5b. GRANT NUMBER		5c. COOPERATIVE AGREEMENT NUMBER
5d. PROGRAM ELEMENT NUMBER 633007	5e. PROJECT NUMBER A792	5f. TASK NUMBER 1029	5g. WORK UNIT NUMBER	
6. AUTHOR(S) Deirdre J. Knapp & Cristina D. Kirkendall				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization 66 Canal Center Plaza, Suite 700 Alexandria, Virginia 22314				8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 6000 6th Street (Bldg. 1464 / Mail Stop: 5610) Fort Belvoir, Virginia 22060-5610		10. SPONSOR/MONITOR'S ACRONYM(S) ARI		11. SPONSOR/MONITOR'S REPORT NUMBER(S) Technical Report 1437
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.				
13. SUPPLEMENTARY NOTES ARI Research POC: Dr. Cristina D. Kirkendall, Selection and Assignment Research Unit				
14. ABSTRACT In addition to educational, physical, and moral screens, the U.S. Army relies on the Armed Forces Qualification Test (AFQT), a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), to select new Soldiers into the Army. Although AFQT is useful for selecting new Soldiers, other personal attributes are important to Soldier performance and retention. Based on previous U.S. Army Research Institute for the Behavioral and Social Sciences investigations, the Army selected one promising measure, the Tailored Adaptive Personality Assessment System (TAPAS), beginning administration to applicants in 2009. Since that time, criterion data were compiled at 6-month intervals from administrative records, schools for selected military occupational specialties (MOS), and Soldiers in units. This is part of a series of biannual evaluations of TAPAS that has been done under a talent management support initiative known as <i>Validation of Accessions Screening Tools</i> (VAST). The cumulative results suggest that TAPAS scores significantly predict a number of criteria of interest (e.g., attrition, Army fit), indicating that the measure holds promise for enlisted personnel selection.				
15. SUBJECT TERMS Behavioral and social science; Personnel; Manpower; Selection and classification				
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified	Unlimited Unclassified	
			18. NUMBER OF PAGES 103	
19a. NAME OF RESPONSIBLE PERSON Tonia S. Heffner			19b. PHONE NUMBER (Include area code) 703-545-4408	

Technical Report 1437

**Exploring Personality and Performance Criterion
Relationships Using the Tailored Adaptive Personality
Assessment System**

Editors

Deirdre J. Knapp

Human Resources Research Organization

Cristina D. Kirkendall

U.S. Army Research Institute

Selection and Assignment Research Unit

Tonia S. Heffner, Chief

July 2023

Approved for public release; distribution is unlimited.

ACKNOWLEDGEMENT

We would like to acknowledge the Army cadre who support criterion data collection efforts at the schoolhouses. These noncommissioned officers (NCOs) ensure that trainees are scheduled to take the research measures and provide ratings of their Soldiers' performance in training. Those Army personnel who support the in-unit data collections are also instrumental to this research program.

We also extend our appreciation to the Army Test Program Advisory Team (ATPAT), a group of senior NCOs who periodically meet with U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) researchers to help guide this work in a manner that ensures its relevance to the Army and who help enable the Army support required to implement the research.

Members of the ATPAT are:

Mr. Shawn Bova
Mr. Alphonsa Green
SGM Latoya Hammond
Mr. James Hayes
SGM Robert Haynie
Mr. Lawrence Hayward
1SG Thomas Hill
MSG Kaulai Hollis
MSG Richard Howe
SFC Saba James
SGM Louis Johnson
MSG Chad Kenney
SGM Gareth Kilpatrick
SFC Clifford Kurten
SGM Alfredo Lassiseramirez

SFC Ricardo Martinez
CSM Shawn McDonald
SGM Jeffrey Morris
CW5 Anthony Moschella
SFC Anthony Northcutt
Mr. John Plotts
Mr. Gerald Purcell
MSG Joel Raglin
MSG Felicia Redd
Mr. Mark Russo
CSM Todd Shirley
SGM Thomas Skelton
SFC Marvin Skinner
1SG David Smith
1SG Christopher Stanifer
Ms. Natalya Weber

EXPLORING PERSONALITY AND PERFORMANCE CRITERION RELATIONSHIPS USING THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM

EXECUTIVE SUMMARY

Research Requirement:

In addition to educational, physical, and moral screens, the U.S. Army relies on the Armed Forces Qualification Test (AFQT), a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), to select new Soldiers into the Army. Although AFQT has proven to be and will continue to serve as a useful metric for selecting new Soldiers, there is a growing recognition of the need to consider whole-person assessment that takes other personal attributes, in particular non-cognitive attributes (e.g., temperament, interests, and values) into consideration. Non-cognitive attributes are important to entry-level Soldier performance and retention (e.g., Campbell & Knapp, 2001; Ingerick, Diaz, & Putka, 2009; Knapp & Heffner, 2009, 2010; Knapp & Tremble, 2007). Based on previous research (Knapp & Heffner, 2010), the Army selected one particularly promising measure, the Tailored Adaptive Personality Assessment System (TAPAS), as the basis for an initial operational test and evaluation (IOT&E) of the *Tier One Performance Screen (TOPS)*. TAPAS capitalizes on the latest advances in testing technology to assess motivation through the measurement of personality characteristics. The TOPS IOT&E evolved into a long-term personnel database effort that supports empirically based policymaking across a wide range of talent accession decisions, including those related to the operational use of TAPAS. This broader-based effort is now known as the *Validation of Accessions Screening Tools (VAST)* research program and it serves multiple needs in support of the Army's People Strategy (U.S. Army, 2019).

Procedure:

In May 2009, the U.S. Army began administering TAPAS at Military Entrance Processing Stations (MEPS). To validate TAPAS, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) began collecting outcome (criterion) data at multiple points in time from Soldiers who took TAPAS at entry. At present, initial military training (IMT) criterion data are being collected at schools for Soldiers in 71 military occupational specialties (MOS). Project teams also collect criterion data from Soldiers in units (regardless of MOS) in annual waves of site visits to Army installations.

Starting with an extract of the Total Army Personnel Database (TAPDB), researchers biannually match ASVAB and TAPAS scores with criterion data collected specifically for ARI as well as data from administrative sources (e.g., course grades, attrition). The ARI criterion measures include job knowledge tests, an attitudinal assessment (the Army Life Questionnaire), and performance rating scales completed by the Soldiers' supervisors (in units) and their peers (in both IMT and in units).

The data presented in this report include TAPAS and criterion data collected through July 2020. The TAPAS database consists of over 1.2M applicants. The validation sample sizes are considerably smaller, with the IMT Validation Sample comprising 103,628 Soldiers, the In-Unit Validation Sample comprising 9,598 Soldiers, and the Administrative Validation Sample (which

includes Soldiers with criterion data [e.g., attrition] from at least one administrative source) comprising 741,809 Soldiers.

Data from job knowledge tests, attitudinal assessment, performance ratings provided by supervisors and/or peers, and administrative sources were combined to yield an array of scores representing important Soldier outcomes. In general, the criterion scores exhibited acceptable psychometric properties. The exception to this conclusion was the performance rating scales which exhibited relatively low inter-rater reliability. As a result, findings involving the rating scale scores may underestimate relationships with other variables.

Our approach to analyzing the incremental predictive validity of TAPAS over AFQT in this report was consistent with previous evaluations of this measure and similar experimental non-cognitive predictors (e.g., Ingerick et al., 2009; Knapp & Heffner, 2009, 2010, 2011). In brief, this approach involved testing a series of hierarchical regression models, regressing scores for each criterion measure onto Soldiers' AFQT scores, followed by their TAPAS composite or TAPAS scale scores in the second step. The resulting increment in the multiple correlation value (ΔR) when TAPAS composite or scale scores were added to the baseline regression models served as our index of incremental validity. Analyses used the TAPAS Will-Do (motivation), Can-Do (knowledge), and Adaptation (attrition) composite scores.

Findings:

Results presented here align with the findings from prior analysis cycles (e.g., Hughes, 2018). Overall, the TAPAS composite scores demonstrated incremental validity over AFQT in predicting several important first-term Soldier outcomes. In regard to prediction of continuous criterion measures, the TAPAS Will-Do composite demonstrated the greatest incremental validity overall, with ΔR estimates exceeding .05 for predicting peer performance ratings, Army Physical Fitness Test (APFT) scores, peer leadership, and resilience for Soldiers in IMT and in units, as well as commitment/fit and counterproductive Soldier behavior (reverse scored) for Soldiers in IMT. The TAPAS Adaptation composite also yielded incremental validity estimates of .05 or greater for APFT scores and resilience for Soldiers in IMT and in units, as well as for peer performance ratings in IMT. Both the Will-Do and Adaptation composites also demonstrated statistically significant incremental relationships beyond AFQT with three dichotomous criteria (i.e., attrition, disciplinary incidents, and IMT restarts). Higher scores on these composites were associated with a reduction in the likelihood of these outcomes. Conversely, the TAPAS Can-Do composite did not provide any notable incremental validity for the criteria examined in this research. This finding is not surprising given the established strength of AFQT to predict the same criteria as the Can-Do composite was designed to predict, as well as the moderate correlation between AFQT and Can-Do composite scores.

As an ancillary analysis, we created TAPAS composite scores to approximate two Big Five personality constructs – Conscientiousness and Emotional Stability (McCrae & Costa, 1987). These scores generally exhibited relationships with performance in the direction of findings observed in the civilian literature, but the relationships were smaller than expected. Several possible reasons for these findings are offered. These analyses used multilevel modeling to detect differences in observed relationships across MOS.

Utilization and Dissemination of Findings:

The VAST research findings will be used by the Assistant Secretary of the Army (Manpower and Reserve Affairs); Army Deputy Chief of Staff, G-1; U.S. Army Training and Doctrine Command; U.S. Army Recruiting Command; and to evaluate the effectiveness of tools used for Army applicant selection and assignment.

EXPLORING PERSONALITY AND PERFORMANCE CRITERION RELATIONSHIPS USING THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM

CONTENTS

	Page
CHAPTER 1: INTRODUCTION	1
Deirdre J. Knapp (HumRRO) and Cristina D. Kirkendall (ARI)	
Background	1
The Tailored Adaptive Personality Assessment System (TAPAS).....	2
The Validation of Accessions Screening Tools (VAST) Program.....	3
Overview of Report	3
CHAPTER 2: RESEARCH METHOD.....	4
Deirdre J. Knapp, Arielle Rogers, Martin Yu (HumRRO), and Chris Nye (DCG)	
Overview	4
Data Collection.....	4
Biannual Longitudinal Database Updates	5
Description of Analysis Samples	6
Predictor Measures	9
CHAPTER 3: VAST PROGRAM CRITERION MEASURES.....	14
Alyssa D. Marshall, Colby K. Nesbitt, Timothy Burgoyne, and Deirdre J. Knapp (HumRRO)	
Objective Criterion Measures.....	16
Army Life Questionnaire.....	20
Performance Rating Criterion Measures	23
Cross-Instrument Correlations and Subgroup Differences	31
Summary	32
CHAPTER 4: INCREMENTAL VALIDITY OF PROPOSED TAPAS COMPOSITE SCORES.....	33
Tetsuhiro Yamada and Michael G. Hughes (HumRRO)	
Bivariate Correlations.....	33
Incremental Validity Analyses	33
Summary	50
CHAPTER 5: EXAMINATION OF SELECTED TEMPERAMENT CONSTRUCTS.....	51
Stephen A. Robertson and Samuel J. Posnock (HumRRO)	
Introduction	51
Method.....	52
Results	55
Discussion	60
Summary	61

CONTENTS (CONTINUED)

	Page
CHAPTER 6: REFLECTIONS.....	62
Deirdre J. Knapp (HumRRO) and Cristina D. Kirkendall (ARI)	
Criterion-Related Validity Evidence	62
Other TAPAS Evaluation Efforts.....	63
What's Next for TAPAS?.....	64
REFERENCES	65

APPENDICES

APPENDIX A - LIST OF FREQUENTLY USED ACRONYMS.....	A-1
APPENDIX B - PREDICTOR MEASURE PSYCHOMETRIC PROPERTIES IN THE APPLICANT SAMPLE.....	B-1
APPENDIX C - CRITERION MEASURE PSYCHOMETRIC PROPERTIES.....	C-1
APPENDIX D - BIVARIATE CORRELATIONS AMONG PREDICTOR AND CRITERION SCORES	D-1

LIST OF TABLES

Table 2.1. Background and Demographic Characteristics of the Analysis Samples.....	7
Table 2.2. TAPAS Versions by Administration Date.....	10
Table 2.3. TAPAS Facet Names and Definitions	11
Table 3.1. IMT Criterion Scores	14
Table 3.2. In-Unit Criterion Scores.....	15
Table 3.3. Attrition Criterion Scores.....	16
Table 3.4. Job Knowledge Test (JKT) Reliability Estimates in the IMT and In-Unit Validation Samples	18
Table 3.5. Job Knowledge Test (JKT) Descriptive Statistics in the IMT and In-Unit Validation Samples	19
Table 3.6. Base Rates for Dichotomous Criteria in the Administrative Validation Sample	20
Table 3.7. Army Life Questionnaire (ALQ) Scale and Composite Reliability Estimates in the IMT and In-Unit Validation Samples	22
Table 3.8 Army Life Questionnaire (ALQ) Descriptive Statistics in the IMT and In-Unit Validation Samples	23

CONTENTS (CONTINUED)

	Page
Table 3.9. IMT PRS-P Rating Scale Dimensions and Associated Example Behaviors	25
Table 3.10. In-Unit PRS-P Rating Scale Dimensions and Associated Example Behaviors	27
Table 3.11. PRS-P Interrater Reliability Estimates in the IMT and In-Unit Validation Samples	28
Table 3.12. Descriptive Statistics for the Peer Performance Rating Scales (PRS-P) in the IMT Validation Sample	29
Table 3.13. Descriptive Statistics for the Army-Wide Performance Rating Scales (PRS-P) in the In-Unit Validation Sample	29
Table 3.14. In-Unit PRS-S Army-Wide Rating Scale Dimensions and Associated Definitions	30
Table 4.1. Incremental Validity Estimates for TAPAS Composites over AFQT for Predicting IMT Criteria	35
Table 4.2. Incremental Validity Estimates for TAPAS Composites over AFQT for Predicting In-Unit Criteria	38
Table 4.3. Incremental Validity Estimates for TAPAS Composites over AFQT for Predicting IMT Restarts and Disciplinary Incidents	42
Table 4.4. Incremental Validity Estimates for TAPAS Composites over AFQT for Predicting Cumulative Attrition through 36 Months of Service (Regular Army Only)	46
Table 5.1 TAPAS Facets and Trait Composites Descriptive Statistics	55
Table 5.2. Correlations Among Predictor Measures	56
Table 5.3. Criterion Scale Descriptive Statistics for Analysis Sample	56
Table 5.4. Criterion Score Intercorrelations	57
Table 5.5 Factor Analysis Results	57
Table 5.6 Level 2 Variance Results for Multilevel Models	58
Table 5.7 IMT Model Results	59
Table 5.8 In-Unit Model Results	60

LIST OF FIGURES

Figure 2.1. Overview of VAST data file merging and nested sample generation process.	6
Figure 3.1. Sample IMT peer rater performance rating scales.	26
Figure 3.2. Sample in-unit supervisor rater performance rating scale.	31
Figure 4.1. Increase in prediction of IMT criteria using TAPAS Can-Do composite.	36
Figure 4.2. Increase in prediction of IMT criteria using TAPAS Will-Do composite.	36
Figure 4.3. Increase in prediction of IMT criteria using TAPAS Adaptation composite.	37
Figure 4.4. Increase in prediction of IMT criteria using relevant TAPAS facets.	37
Figure 4.5. Increase in prediction of in-unit criteria using TAPAS Can-Do composite.	39
Figure 4.6. Increase in prediction of in-unit criteria using TAPAS Will-Do composite.	40
Figure 4.7. Increase in prediction of in-unit criteria using TAPAS Adaptation composite.	40
Figure 4.8. Increase in prediction of in-unit criteria using relevant TAPAS facets.	41
Figure 4.9. Predictive accuracy of AFQT and TAPAS Can-Do composite in the discrimination of both IMT and in-unit Disciplinary Incidents and IMT Restarts.	44
Figure 4.10. Predictive accuracy of AFQT and TAPAS Will-Do composite in the discrimination of both IMT and in-unit Disciplinary Incidents and IMT Restarts.	44
Figure 4.11. Predictive accuracy of AFQT and TAPAS Adaptation composite in the discrimination of both IMT and in-unit Disciplinary Incidents and IMT Restarts.	45
Figure 4.12. Predictive accuracy of AFQT and TAPAS Will-Do facets in the discrimination of both IMT and in-unit Disciplinary Incidents and IMT Restarts.	45
Figure 4.13. Predictive accuracy of AFQT and TAPAS Can-Do composite in the discrimination of Cumulative Attrition.	48
Figure 4.14. Predictive accuracy of AFQT and TAPAS Will-Do composite in the discrimination of Cumulative Attrition.	48
Figure 4.15. Predictive accuracy of AFQT and TAPAS Adaptation composite in the discrimination of Cumulative Attrition.	49
Figure 4.16. Predictive accuracy of AFQT and TAPAS Adaptation facets in the discrimination of Cumulative Attrition.	49

EXPLORING PERSONALITY AND PERFORMANCE CRITERION RELATIONSHIPS USING THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM

CHAPTER 1: INTRODUCTION

Deirdre J. Knapp (HumRRO) and Cristina D. Kirkendall (ARI)

Background

The U.S. Army's People Strategy is geared to ensuring a ready, professional, diverse, and integrated force (U.S. Army, 2019). The strategy envisions a variety of ways to address these goals, including "organizational research that employs scientific methods to develop and deploy leading edge approaches to solving Army people challenges" (p. 9). The Selection and Assignment Research Unit (SARU) of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) has been conducting such research for 50 years. Over the past decade, ARI has investigated the Tailored Adaptive Personality Assessment System (TAPAS) as a tool to allow the Army to have a more *whole person* approach to selecting enlisted personnel and classifying them into Army occupations. Indeed, the latest People Strategy implementation plan (U.S. Army, 2020) specifically calls for using TAPAS scores in two pilot programs to support recruiting goals by expanding the definition of high-quality prospective enlistees. Further research on TAPAS is the focus of the present report. We begin, however, with providing further contextual background.

In addition to educational, physical, and moral screens, the U.S. Army relies on the Armed Forces Qualification Test (AFQT), a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), to select new Soldiers into the Army. Although AFQT has proven to be, and will continue to serve as, a useful metric for selecting new Soldiers, other personal attributes, in particular non-cognitive attributes (e.g., temperament, interests, and values), are important to entry-level Soldier performance and retention (e.g., Knapp & Tremble, 2007). Importantly, broadening the lens through which potential new recruits are considered not only could result in better decision-making, it is also critical for helping to ensure a demographically diverse enlisted force. That said, the historical obstacle to using information about such things as temperament and interests in a large-scale personnel system has been measuring these characteristics in a manner that prevents applicants from distorting their responses to meet standards.

The ASVAB has undergone improvements over the years, but the basic content has remained largely unchanged since 1976 when it was initially introduced as the selection and classification battery to be used by all the U.S. military services. Potential content to incorporate into or be used in conjunction with ASVAB has included personality measures (e.g., TAPAS, Drasgow et al., 2012; Navy Computer Adaptive Personality Scales, Houston, Borman, Farmer, & Bearden, 2006), the Cyber Test (Trippe, Moriarty, Russell, Carretta, & Beatty, 2014), interest measures such as the Army Vocational Interest Diagnostic (AVID; Kirkendall et al., 2020) and the Job Opportunities in the Navy (JOIN; Farmer, Watson, Alderton, Michael, & Hindelang, 2006), and improved physical fitness measures (Sharp et al., 2018). An initiative informally known as ASVAB 2.0 is also underway to take a systematic look at all ASVAB content to reassess what

would be of most value to include to support enlisted talent management needs across the U.S. military services.

The Tailored Adaptive Personality Assessment System (TAPAS)

ARI's research on the TAPAS dates back to approximately 2008 and was preceded by earlier lines of research seeking ways to improve the Army's selection and classification systems by expanding the psychological assessment information on which these systems are based (i.e., cognitive ability) and broadening the consideration of outcomes of interest (e.g., to include motivational aspects of job performance). Related prior ARI research programs included the seminal Project A conducted in the 1980s through early 1990s (Campbell & Knapp, 2001) and the Select21 and Army Class programs conducted in the early 2000s (Allen, Knapp, & Owens, 2016; Knapp, Sager, & Tremble, 2005; Knapp & Tremble, 2007).

Validating Future Force Performance Measures (Army Class) examined the potential of several experimental non-cognitive measures (e.g., temperament and person-environment fit) developed in the Select21 project for predicting important Army outcomes (e.g., performance, attitudes, attrition). The Army Class research project was a six-year effort conducted with contract support from the Human Resources Research Organization (HumRRO; Allen et al., 2016; Ingerick, Diaz, & Putka, 2009; Knapp & Heffner, 2009). In a contemporaneous but separate effort, the TAPAS was developed by the Drasgow Consulting Group (DCG) under a Small Business Innovative Research program (Drasgow et al, 2012). Given the promise of the TAPAS for addressing applicant faking concerns by using the latest advances in psychometric theory, it was integrated into late stages of the Army Class validation effort and surfaced as one of the most promising measures investigated in that research (Knapp & Heffner, 2010).

Based on this research, the Army requested that TAPAS be administered to Army applicants at Military Entrance Processing Stations (MEPS) on the computer adaptive platform for the ASVAB (CAT-ASVAB). Testing of Education Tier 1 (primarily high school diploma graduate) applicants began in May 2009 and was subsequently expanded to all Army applicants. In an associated research program dubbed the *Tier One Performance Screen (TOPS) initial operational test and evaluation (IOT&E)*, ARI investigated the potential for TAPAS to identify Tier 1 applicants who would likely perform differently (higher or lower) than would be predicted by their ASVAB scores. The initial conceptualization for the IOT&E was to use TAPAS as a tool for "screening in" applicants with lower AFQT scores, allowing more of these applicants to access into the Army (hence the research program title). However, changing economic conditions could spur use of TAPAS as a tool to screen out low-motivated applicants, thus making the selection criteria to enter the Army more stringent. As this change in strategy illustrates, decision rules regarding the derivation and use of TAPAS composite scores could be adjusted in myriad ways to reflect evolving conditions.

ARI's research findings prompted the U.S. Air Force, Navy, and Marine Corps to begin thinking about how they might want to use TAPAS for selection (e.g., Trent, Barron, Rose, & Carretta, 2020) and classification (e.g., Rose, Manley, & Weissmuller, 2013;). Increasing interest also prompted a comprehensive evaluation of TAPAS sponsored by the Defense Personnel Assessment Center (DPAC) which was completed in 2019 (Roberts, Arthur, Reckase, Sackett, & Zenisky, 2019).

The Validation of Accessions Screening Tools (VAST) Program

Generation of criterion-related validity evidence for using TAPAS for Army enlisted personnel selection, which initially began as part of the TOPS IOT&E project, is being continued in the current research effort, *Validation of Accessions Screening Tools (VAST)*. While the method of conducting ongoing data collections and biannual criterion-related validation analyses is the same as it was under the TOPS IOT&E, the resulting VAST database is expressly intended to support myriad research efforts pertinent to Army enlisted talent acquisition and management needs including, but not limited to, investigations related to TAPAS.

Under VAST, ARI collects training outcome data on Soldiers in 71 military occupational specialties (MOS) as they complete initial military training (IMT). The criterion (outcome) measures include job knowledge tests, an attitudinal assessment, and peer performance ratings. These measures are computer-administered in a process overseen by Army personnel with guidance and support from both ARI and HumRRO.

Criterion data are also collected from Soldiers and their supervisors in units during annual data collections at major Army installations. These proctored “in-unit” data collections began in January 2011 and target Soldiers who accessed into the Army since approximately 2010. The in-unit criterion measures parallel those administered in IMT, but also include supervisor performance ratings. Separation status and other administrative data for all Soldiers is updated biannually.

This report is the third analysis cycle under VAST but serves as the 21st iteration of the TAPAS criterion-related validation analyses that began with the TOPS IOT&E. Prior ARI TAPAS biannual validation efforts are described in a series of technical reports (Bynum & Mullins, 2017; Knapp & Heffner, 2011, 2012; Knapp, Heffner, & White, 2011; Knapp & Kirkendall, 2018, 2020; Knapp & LaPort, 2013a, 2013b, 2014; Knapp & Wolters, 2017).

To date, both the TOPS IOT&E and VAST research has focused on consideration of the incremental validity of TAPAS composite scores over the use of AFQT scores for selection. It is important to recognize that these analyses do not factor in other qualification requirements which include moral and physical fitness screens.

Overview of Report

Chapter 2 describes the VAST research method, including the predictor (ASVAB and TAPAS) measures and their associated psychometric properties, criterion data collection procedures, and database construction strategy. It closes by describing the Cycle 21 analysis samples in detail. Chapter 3 describes the criterion measures and their associated psychometric properties. Chapter 4 describes criterion-related validation analyses, focusing on the incremental validity of three TAPAS composite scores over AFQT for predicting key outcomes. In contrast to the operational system focus of Chapter 4, Chapter 5 looks at the predictive value of two *Big Five* personality constructs measured by the TAPAS (conscientiousness and emotional stability) and compares these findings to those seen in the civilian research literature. The report concludes with Chapter 6 which reflects on our research findings and looks ahead to future research.

CHAPTER 2: RESEARCH METHOD

Deirdre J. Knapp, Arielle Rogers, Martin Yu (HumRRO), and Chris Nye (DCG)

Overview

The VAST research uses a longitudinal criterion-related validation evidence design. Army applicants take ASVAB and TAPAS as part of the usual pre-enlistment accessioning process. We track all those who enlist by, at a minimum, retrieving available administrative outcome data (e.g., separation status) from Army records. We also seek to collect for-research-only criterion data from many of these Soldiers as they complete IMT by having ongoing data collection activity at a large number of MOS schoolhouses. Criterion data are collected from Soldiers who have successfully moved on to their assigned units through annual visits to various Army installations. We refer to these as the IMT and in-unit data collections.

Twice per year, we update the VAST longitudinal database with newly collected IMT and in-unit data, as well as updated administrative data. After each new iteration of the database becomes available, we analyze the data and document the results. Carrying forward the nomenclature from the TOPS IOT&E research, results in this report are based on the 21st cycle of biannual database updates.

Criterion outcomes include attitudinal measures, job knowledge test scores, peer and supervisor performance ratings, and attrition. These measures are described further in Chapter 3. The remainder of this chapter provides details of the criterion data collection method, pre-enlistment predictor measures, database construction process, and characteristics of the latest analysis samples.

Data Collection

IMT Data Collections

Cadre at participating schoolhouses are generally responsible for ensuring that Soldiers take the online VAST IMT criterion assessment approximately 2 weeks prior to graduation. The goal is to schedule the assessment toward the end of training, but before the intense period of activity right before Soldiers graduate.

After completing a job knowledge test and attitudinal measure, Soldiers in IMT are directed by instructional screens to find their name on a copy of the class roster, then provide performance ratings for the five Soldiers who follow them on the roster. Use of the rosters makes the peer rater identification process very simple, while also providing Soldiers with the correct spelling of the rated peers' names. Although cadre provided performance ratings for their Soldiers in the past, we eliminated this requirement several years ago to minimize burden on the schools.

In-Unit Data Collections

Most large Army Forces Command installations host an annual “umbrella week” which are time periods set aside to support research data collection activities. We schedule data collection visits during these umbrella weeks, resulting in rolling data collection activity throughout the year at various locations. Soldiers and their first-line supervisors are instructed to report to a computer classroom at which time data collection staff (ARI and/or HumRRO researchers) guide them through the online data collection process.

To identify peer rater-ratee pairs in units, data collectors ask supervisors arriving to the test site to organize their Soldiers into groups according to who each supervisor can rate. As each group signs in, data collectors ask the Soldiers, as a group, if they have worked with each other long enough to be familiar with each other’s performance. If the group includes more than six Soldiers, only five peers are assigned to each Soldier. Soldiers who are determined to have no peers present are instructed to skip the peer ratings section of the assessment.

Biannual Longitudinal Database Updates

Twice per year, the VAST longitudinal database is updated with new cases and updated data for Soldiers who were introduced into the database in earlier cycles.

An illustrative view of our data sources is provided in Figure 2.1 (see Appendix A for acronym definitions). The analysis data files are built on a foundational database drawn from the Total Army Personnel Database (TAPDB) onto which predictor and criterion data from other sources are added. The lighter boxes within the figure represent sources of data, and the darker boxes represent samples on which descriptive or inferential analyses are conducted. The leftmost column in the figure summarizes the predictor and primary demographic data sources used to derive the *Applicant Sample*. The other columns summarize the research-only (i.e., non-administrative) and administrative criterion data that are also matched to TAPDB records. Soldiers with both predictor and criterion data form the *IMT*, *In-Unit*, and *Administrative Validation Samples*.

For the purpose of evaluating the effectiveness of TAPAS, exclusions to the analysis samples were imposed based on prior military service, AFQT score, and education level. Specifically, applicants who previously served in the military are dropped. AFQT Category V cases are excluded because they are not eligible for military service. Similarly, non-high school graduates (Education Tier 3 – individuals not currently attending high school and do not possess a high school diploma or alternate credential) are removed because they are also ineligible for accession.

The Applicant Sample includes Soldiers from all Army components – Regular Army (RA), U.S. Army Reserve (USAR), and U.S. Army National Guard (ARNG) who took TAPAS prior to enlistment. Soldiers from all components are included in all analyses except for those involving attrition. For those analyses, only Regular Army Soldiers are included.

It bears noting that the TOPS IOT&E research databases (Cycles 1 through 18) used the TAPAS database provided by Military Entrance Processing Command (MEPCOM) as the foundational database rather than TAPDB. Using the much larger TAPDB makes the foundational database more useful to ARI for a variety of personnel research needs.

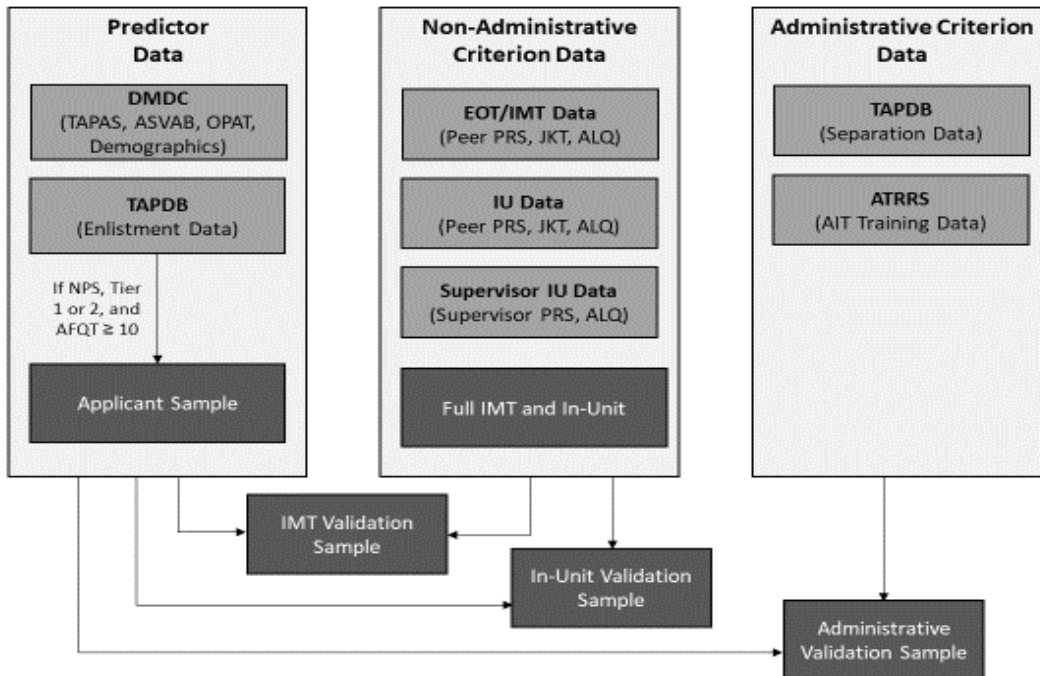


Figure 2.1. Overview of VAST data file merging and nested sample generation process.

We begin by describing how Soldier data are filtered for analysis, then describe multiple subsamples that are created to support various types of analyses, and, finally, provide descriptive statistics for the predictor variables.

Description of Analysis Samples

Table 2.1 summarizes the key variables that were used to create the analysis samples and the demographic characteristics observed in those samples. The Applicant Sample includes Soldiers:

- in Education Tier 1 or 2 (high school diploma or non-diploma graduates),
- without prior service history,
- in AFQT Category I-IV (eligible for military service),
- with valid TAPAS score data (e.g., without flat responding, large amounts of missing data), and
- who were administered TAPAS versions other than the 13D-CAT.

After applying these filters, 1,261,864 Soldiers met the screening criteria and were included in the Cycle 21 Applicant Sample.

Table 2.1. Background and Demographic Characteristics of the Analysis Samples

Characteristic	Applicant ^a		Administrative Validation ^b		IMT Validation ^c		In-Unit Validation ^d	
	<i>n</i> = 1,261,864		<i>n</i> = 741,809		<i>n</i> = 103,628		<i>n</i> = 9,598	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
<i>Component</i>								
Regular	724,919	57.4	446,610	60.2	62,461	60.3	9,523	99.2
ARNG	372,647	29.5	202,955	27.4	30,047	29.0	42	0.4
USAR	164,298	13.0	92,244	12.4	11,120	10.7	33	0.3
<i>Education Tier</i>								
Tier 1	1,199,195	95.0	711,199	95.9	99,650	96.2	9,294	96.8
Tier 2	62,669	5.0	30,610	4.1	3,978	3.8	304	3.2
<i>Military Occupational Specialty</i>								
Infantryman (11B/C/X + 18X)	127,835	10.1	109,554	14.8	20,154	19.4	2,204	23.0
Combat Engineer (12B)	24,420	1.9	20,542	2.8	6,969	6.7	379	3.9
Cannon Crewmember (13B)	16,386	1.3	13,488	1.8	576	0.6	204	2.1
Field Artillery Tactical Data Systems Specialist (13D)	4,533	0.4	4,419	0.6	984	0.9	65	0.7
Fire Support Specialist (13F)	9,268	0.7	7,723	1.0	1,202	1.2	76	0.8
Cavalry Scout (19D)	21,741	1.7	18,394	2.5	2,840	2.7	464	4.8
M1 Armor Crewmember (19K)	9,914	0.8	8,690	1.2	2,703	2.6	257	2.7
Military Police (31B)	35,499	2.8	31,860	4.3	12,936	12.5	169	1.8
Human Resources Specialist (42A)	21,856	1.7	19,106	2.6	4,604	4.4	242	2.5
Combat Medic Specialist (68W)	39,889	3.2	33,297	4.5	9,016	8.7	366	3.8
Motor Transport Operator Soldiers (88M)	40,614	3.2	35,235	4.7	12,302	11.9	668	7.0
Wheeled Vehicle Mechanic (91B)	42,911	3.4	35,957	4.8	3,371	3.3	525	5.5
Other	445,688	35.3	380,113	51.2	25,970	25.1	3,979	41.5
Unknown	421,310	33.4	23,431	3.2	1	0.0	--	--
<i>AFQT Category^e</i>								
I	71,480	5.7	43,220	5.8	6,331	6.1	408	4.3
II	360,304	28.6	230,502	31.1	35,528	34.3	2,518	26.2
IIIA	256,794	20.4	164,650	22.2	23,751	22.9	2,151	22.4
IIIB	392,992	31.1	258,522	34.9	31,809	30.7	3,851	40.1
IV	180,294	14.3	44,915	6.1	6,209	6.0	670	7.0

Table 2.1. (Continued)

Characteristic	Applicant ^a		Administrative Validation ^b		IMT Validation ^c		In-Unit Validation ^d	
	<i>n</i> = 1,261,864		<i>n</i> = 741,809		<i>n</i> = 103,628		<i>n</i> = 9,598	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
<i>Contract Status</i>								
Signed	879,209	69.7	--	--	--	--	--	--
Not signed	382,655	30.3	--	--	--	--	--	--
<i>Gender</i>								
Female	292,666	23.2	147,455	19.9	17,110	16.5	1,191	12.4
Male	969,198	76.8	594,354	80.1	86,518	83.5	8407	87.6
<i>Race</i>								
African American	298,322	23.6	161,479	21.8	18,562	17.9	2,298	23.9
American Indian	10,166	0.8	5,811	0.8	813	0.8	72	0.8
Asian	59,661	4.7	34,003	4.6	4,479	4.3	551	5.7
Hawaiian/Pacific Islander	2,694	0.2	1,380	0.2	148	0.1	23	0.2
Caucasian	873,519	69.2	531,588	71.7	78,659	75.9	6,565	68.4
Multiple	3,920	0.3	2,287	0.3	242	0.2	32	0.3
Declined to Answer/Missing	13,582	1.1	5,261	0.7	725	0.7	57	0.6
<i>Ethnicity</i>								
Hispanic/Latino	213,369	16.9	118,655	16	18,087	17.5	1,815	18.9
Not Hispanic	1,035,2	82.0	618,465	83.4	84,924	82.0	7,721	80.4
Declined to Answer/Missing	13,295	1.1	4,689	0.6	617	0.6	62	0.6

^aNon-prior service Soldiers in Education Tier 1 and 2 and AFQT Category IV or above, with valid TAPAS score data on forms other than the 13D-CAT.

^bSoldiers in Applicant Sample with a signed contract and at least one criterion record (i.e., schoolhouse, in-unit, ATRRS, RITMS, or attrition).

^cSoldiers in Applicant Sample with a signed contract and criterion data collected at schoolhouses.

^dSoldiers in Applicant Sample with a signed contract and criterion data collected in units.

^eAFQT Categories IIIB and IV are oversampled. Values are not representative of Army accessions.

The Administrative Validation Sample includes 741,809 Soldiers who met all the inclusion criteria for the Applicant Sample and also had a signed contract and at least one record from an administrative criterion data source (e.g., Army Training Requirements and Resources System [ATRRS]). There are 103,628 Soldiers with IMT criterion data linked to an administrative TAPAS record and included in the IMT Validation Sample. There are 9,598 Soldiers with in-unit data and a matching TAPAS record included in the In-Unit Validation Sample.

The Applicant Sample is much larger than the Administrative Validation Sample because not all applicants access into the Army. Similarly, the IMT and In-Unit Validation Samples are much smaller than the Administrative Validation Sample largely because not all Soldiers participate in IMT and in-unit data collections. Furthermore, we rely on self-reported name and date of birth to match TAPAS records to the IMT and in-unit criterion data, which results in a nontrivial percentage of unsuccessful matches.

Analysis sample sizes are generally smaller than the figures reported here because of further data filtering or disaggregation that occurs during analysis. For example, criterion scores are

determined to be valid if they pass multiple data quality screens intended to identify unmotivated responding. Additional screens are analysis-specific and thus not applied to the descriptive analysis of the samples described in this chapter.

Predictor Measures

Armed Forces Qualification Test

The ASVAB is a multiple aptitude battery of tests administered by MEPCOM. Most military applicants take the computer-adaptive version of ASVAB. Scores on the ASVAB tests are combined to create composite scores for use in selecting applicants into the Army and qualifying them for various MOS. AFQT, the composite used for selecting applicants into the Army, comprises Verbal Expression (VE), Arithmetic Reasoning (AR), and Math Knowledge (MK) scores ($AFQT = 2*VE + AR + MK$). It is designed to be a measure of general cognitive ability (Maier, 1993) and has strong psychometric properties in terms of reliability and criterion-related validity (e.g., Campbell & Knapp, 2001; Carretta, Ree, & Teachout, 2016; DMDC, 2008; Hunter, 1986; Thacker, Russell, Knapp, & Gribben, 2020).

AFQT percentile scores are divided into the following categories:

- Category I (93-99)
- Category II (65-92)
- Category IIIA (50-64)
- Category IIIB (31-49)
- Category IV (10-30)
- Category V (1-9)

Per Department of Defense (DoD) regulations, AFQT Category V Soldiers are not eligible for enlistment. Category IV accessions are greatly restricted, some restrictions are placed on Category IIIB accessions, and priority is given to Category I-IIIA accessions.

Tailored Adaptive Personality Assessment System (TAPAS)

TAPAS consists of pools of statements descriptive of various facets of temperament (Drasgow et al., 2012). Items are created by pairing statements of comparable social desirability and extremity (valence) from different facet pools. Statements are selected and paired adaptively to triangulate on estimates of an examinee's standing on the various facets.

As summarized in Table 2.2, there have been multiple versions of TAPAS. The versions vary in which of over two dozen facets of personality are measured. The different versions have allowed ARI to explore the value of alternative facets and to retire the original statement pools that were exposed in research settings. All versions starting with 15D-CAT v5 use DoD-owned statement pools. In the present report, we use data from all TAPAS versions except for the first one (13D-CAT) in which some of the facets have minor conceptual dissimilarities from the same facets in later versions.

Table 2.2. TAPAS Versions by Administration Date

TAPAS Version	Dates Administered	# of Facets	Adaptive	# of Items
13D-CAT	May 4, 2009 to July 10, 2009	13	Yes	104
15D-Static	July 2009 to October 2009	15	No	120
v4	October 2009 to August 2011	15	Yes	120
v5, v7, v8	August 2011 to September 2013	15	Yes	120
v9, v10, v11	September 2013 – October 2019	13	Yes	120
v12, v13, v14	November 2019 – present	13	Yes	120

The three versions of TAPAS administered to Army applicants from 2013 to 2019, and for which we have by far the most data, each assessed 13 facets, 10 of which were common to all three versions. Since 2019, three new versions of the TAPAS have been administered to Army applicants. Each of these TAPAS versions assess a common set of 13 facets (described in Table 2.3) plus four additional research-only facets that are unique to each version. The facets assessed in each version of the TAPAS were selected with a goal of maximizing the validity of this assessment for predicting several Army outcomes. To do so, DCG researchers used data from previous VAST data collections to identify the TAPAS facets that were the best predictors of these criteria. The 13 facets that provided the best prediction are now being assessed in TAPAS v12, v13, and v14. In contrast, the four additional research-only facets that are unique to each version were included to collect operational data on those facets to help determine their future utility.

Scores are standardized within TAPAS versions to enable cross-version analyses. Versions v12, v13, and v14 are set up such that the core 13 facets are administered separately from the experimental mix of facets, thus making the three versions identical for analysis purposes.

The number and composition of TAPAS composites used by the Army have evolved over the years. In all cases, composites have been constructed to predict specific outcomes of interest. While detailed information about the composites is confidential, the changes can be roughly summarized as follows:

- Unit-weighted Can-Do and Will-Do composites were created in 2009 (Knapp & Heffner, 2010) for the early original TAPAS versions.
- Three regression-weighted composites (Can-Do, Will-Do, and Adaptation) were introduced in 2011 (Nye et al., 2012) with the second generation of TAPAS versions (v4-v8). Adaptation was designed to predict early-term attrition.
- Composition (regression weights and/or constituent facets) of the three composites was updated in 2013 (Nye et al. 2014) with the third generation of TAPAS versions (v9-v11).
- Composition of the three composites was updated in 2019 with the latest generation of TAPAS versions (v12-v14), and an additional composite (Conduct) was added to predict attrition due to misbehavior.

Table 2.3. TAPAS Facet Names and Definitions

Facet Name	Brief Description
Achievement	High scoring individuals are seen as hard working, ambitious, confident, and resourceful.
Adjustment	High scoring individuals are well-adjusted, worry-free, and handle stress well.
Attention Seeking	High scoring individuals tend to engage in behaviors that attract social attention. They are loud, loquacious, entertaining, and even boastful.
Dominance	High scoring individuals are domineering, “take charge” and are often referred to by their peers as “natural leaders.”
Even Tempered	High scoring individuals tend to be calm and stable. They don’t often exhibit anger, hostility, or aggression.
Intellectual Efficiency	High scoring individuals believe they process information and make decisions quickly; they see themselves (and they may be perceived by others) as knowledgeable, astute, or intellectual.
Non-Delinquency	High scoring individuals tend to comply with rules, customs, norms, and expectations, and they tend not to challenge authority.
Optimism	High scoring individuals have a positive outlook on life and tend to experience joy and a sense of well-being.
Order	High scoring individuals tend to organize tasks and activities and desire to maintain neat and clean surroundings.
Physical Conditioning	High scoring individuals tend to engage in activities to maintain their physical fitness and are more likely participate in vigorous sports or exercise.
Sociability	High scoring individuals tend to seek out and initiate social interactions.
Team Orientation	High scoring individuals prefer working in teams and make people work together better.
Tolerance	High scoring individuals are interested in other cultures and opinions that may differ from their own. They are willing to adapt to novel environments and situations.

Up until the latest set of revisions, it was generally possible to apply new composite formulae to preceding TAPAS versions for most of the composites. However, none of the TAPAS versions prior to the latest versions measure all the facets needed to compute the three reformulated composite scores. Accordingly, until there are sufficient cases for these composite scores that can be computed using the new formulations, our TAPAS-related analyses will need to treat the old and new composite formulations as comparable. Moreover, since the new Conduct composite score can only be calculated for those who have taken the new TAPAS versions and one of the older versions (v8), there are too few cases with associated criterion data to include Conduct in the validation analyses reported in Chapter 4.

Given the forced-choice nature of the TAPAS response format, traditional methods for estimating reliability of measurement are not suitable. Therefore, alternative estimates of

reliability have been examined for the TAPAS. For example, Zhang et al. (2020) examined the IRT-based marginal reliability estimates for a static version of the TAPAS. Across 10 TAPAS facets, these estimates ranged from .66 to .86. Alternatively, test-retest reliability estimates ranged from .69 to .83 in this same study. Similarly, additional research (Wolters et al., 2017) confirmed these results with several additional TAPAS facets. In two separate samples, the test-retest reliability estimates ranged from .69 to .83. Importantly, this research also examined the test-retest reliabilities of the TAPAS composites and found estimates ranging from .76 to .78. Although these studies suggest that the TAPAS has moderate reliability estimates, more research is needed. All of the studies described above were with civilian samples, using static versions of the TAPAS (i.e., rather than a computer-adaptive version), and conducted in low-stakes (i.e., non-operational) settings. In contrast, studies with military samples have found somewhat lower test-retest reliabilities with estimates ranging from .49 to .74 (average of .61) across eight TAPAS facets and .56, .74, and .70, respectively, for the Can-Do, Will-Do, and Adaptation composites (Wolters et al., 2017). Therefore, more research is needed to evaluate the reliability of the TAPAS in military samples and under operational conditions.

Predictor Score Psychometric Properties

We computed descriptive statistics for AFQT for the entire Applicant Sample ($n = 1,227,106$, $M = 54.55$, $SD = 22.49$). Means and standard deviations for the 13 TAPAS facet scores listed in Table 2.3 and the composite scores are provided in Appendix B (Table B.1). Table 2.4 shows correlations among AFQT and TAPAS composite scores. Appendix B (Table B.2) shows intercorrelations among AFQT and TAPAS facet scores.

Table 2.4. Correlations among the AFQT and TAPAS Composites Scores in the Applicant Sample

Scale	1	2	3	4
1 AFQT				
<i>TAPAS Composites</i>				
2 Can-Do	0.37			
3 Will-Do	0.10	0.15		
4 Adaptation	0.16	0.37	0.55	
5 Conduct	0.06	0.26	-0.08	0.38

Note. Results are limited to the Applicant Sample (which excludes Education Tier 3 and AFQT Category V) with valid TAPAS score data on forms other than the original 13D-CAT. $n = 34,184 - 1,261,290$ All correlations are statistically significant, $p < .01$ (two-tailed).

As might be expected, the Can-Do composite is mostly highly correlated with AFQT ($r = .37$) and weakly correlated with the other three TAPAS composites. The largest composite intercorrelation is between the Will-Do and Adaptation composites ($r = .55$).

Appendix B also shows subgroup differences by gender (Table B.3) as well as race and ethnicity (Table B.4). Effect sizes show statistically significant but relatively small subgroup differences for gender in favor of males for AFQT (-.29), Can-Do (-.27), and Will-Do (-.20) scores. The effect size was somewhat higher (-.40) for the TAPAS Adaptation composite. Females show a small advantage on the Conduct score (.20). Consistent with typical findings related to general

cognitive ability testing, AFQT shows a fairly large effect size in favor of White, Non-Hispanic applicants compared to Black (-.77) and Hispanic applicants (-.58). TAPAS composite scores showed lower effect sizes, ranging from -.01 to -.39 across the three comparison groups. The largest effect sizes tended to be associated with the comparison of White to Asian applicants. The Conduct composite showed minimal evidence of subgroup differences.

CHAPTER 3: VAST PROGRAM CRITERION MEASURES

Alyssa D. Marshall, Colby K. Nesbitt, Timothy Burgoyne, and Deirdre J. Knapp (HumRRO)

This chapter describes the criterion measures used as indicators of outcomes important to the Army (summarized in Tables 3.1-3.3). We describe three categories of criterion measures: (a) objective criteria including those related to the ability of Soldiers to perform their jobs and Soldier attrition; (b) self-reported attitudes and administrative data measured by the Army Life Questionnaire (ALQ); and (c) performance ratings as measured by the Performance Rating Scales for peers (PRS-P) and supervisors (PRS-S), respectively. For each category of outcomes, we provide a historical summary and descriptive statistics.

Table 3.1. IMT Criterion Scores

Criterion	Associated Scales	Description
Peer Performance Ratings: Overall Performance Composite	PRS-P Effort and Discipline PRS-P Physical Fitness PRS-P MOS Qualification Knowledge & Skill PRS-P Resilience and Adjustment PRS-P Working Effectively with Others PRS-P Going Above and Beyond PRS-P Overall Performance Rating	General effort/motivation criterion as rated by peers. Scales are averaged to form the composite.
Commitment and Fit	ALQ Affective Commitment ALQ MOS Fit ALQ Army Fit	General commitment to and fit with the Army. Scales are averaged to form the composite.
Retention Cognitions	ALQ Career Intentions ALQ Reenlistment Intentions	General intentions of continuance in the Army. Scales are averaged to form the composite.
Peer Leadership	ALQ Motivation to Lead-Affective ALQ Motivation to Lead-Noncalculative ALQ Motivation to Lead-Social-Normative ALQ Organizational Citizenship Behavior/Leadership	General aptitude for effective peer leadership in the Army. Scales are averaged to form the composite.
Counterproductive Soldier Behavior	---	ALQ: Intentional behaviors that harm or are intended to harm another Soldier or the legitimate interests of the unit
Resilience	---	ALQ: Capacity to overcome difficult life events with minimal disruption or long-term negative impacts
Knowledge and Skill	WTBD JKT MOS JKT	WTBD JKT and MOS JKT are averaged to form an overall knowledge/skill composite.

Table 3.1. (Continued)

Criterion	Associated Scales	Description
-----------	-------------------	-------------

WTBD JKT	---	Warrior Tasks and Battle Drills (Army-wide) job knowledge test
APFT Score	—	ALQ: Most recent Army Physical Fitness Test Score.
Disciplinary Incidents	—	ALQ: Dichotomous indicator of whether there is at least one “yes” response to a list of possible disciplinary incidents
IMT Restarts	—	Administrative: Indicates whether a Soldier restarted IMT at least once.

Table 3.2. In-Unit Criterion Scores

Criterion	Associated Scales	Description
Supervisor Performance Ratings: Overall Performance Composite	PRS-S Performing Core Warrior Tasks PRS-S Performing MOS-Specific Tasks PRS-S Communicating with Others PRS-S Processing Information PRS-S Solving Problems PRS-S Exhibiting Effort PRS-S Exhibiting Personal Discipline PRS-S Contributing to the Team PRS-S Exhibiting Fitness & Bearing PRS-S Adjusting to the Army PRS-S Following Safety Procedures PRS-S Developing Own Skills PRS-S Managing Personal Matters	General effort/motivation criterion as rated by supervisors. Scales are averaged to form the composite.
Peer Performance Ratings: Overall Performance Composite	PRS-P Effort and Discipline PRS-P Physical Fitness PRS-P MOS Qualification Knowledge & Skill PRS-P Processing Information & Solving Problems PRS-P Resilience and Adjustment PRS-P Working Effectively with Others PRS-P Going Above and Beyond PRS-P Emergent Leadership PRS-P Overall Performance Rating	General effort/motivation criterion as rated by peers. Scales are averaged to form the composite.
Commitment and Fit	ALQ Affective Commitment ALQ MOS Fit ALQ Army Fit	General commitment to and fit with the Army. Scales are averaged to form the composite.
Retention Cognitions	ALQ Career Intentions ALQ Reenlistment Intentions	General intentions of continuance in the Army. Scales are averaged to form the composite.
Peer Leadership	ALQ Motivation to Lead-Affective ALQ Motivation to Lead-Noncalculative ALQ Motivation to Lead-Social-Normative ALQ Organizational Citizenship Behavior/Leadership	General aptitude for effective peer leadership in the Army. Scales are averaged to form the composite.

Table 3.2. (Continued)

Criterion	Associated Scales	Description
-----------	-------------------	-------------

Peer Leadership	ALQ Motivation to Lead-Affective ALQ Motivation to Lead-Noncalculative ALQ Motivation to Lead-Social-Normative ALQ Organizational Citizenship Behavior/Leadership	General aptitude for effective peer leadership in the Army. Scales are averaged to form the composite.
Counterproductive Soldier Behavior	---	ALQ: Intentional behaviors that harm or are intended to harm another Soldier or the legitimate interests of the unit.
Resilience	---	ALQ: Capacity to overcome difficult life events with minimal disruption or long-term negative impacts.
WTBD JKT	---	Warrior Tasks and Battle Drills (Army-wide) job knowledge test
APFT Score	—	ALQ: Most recent Army Physical Fitness Test Score.
Disciplinary Incidents	—	ALQ: Dichotomous indicator of whether there is at least one “yes” response to a list of possible disciplinary incidents.

Table 3.3. Attrition Criterion Scores

Criterion	Description
6-month	Attrition at or before 6 months of service
12-month	Attrition at or before 12 months of service
24-month	Attrition at or before 24 months of service

Objective Criterion Measures

Objective criterion measures are used as indicators of two outcomes of critical importance to the Army: the ability of Soldiers to perform their jobs and reduced Soldier attrition. We assess Soldier performance primarily via job knowledge tests (JKTs). According to Campbell, McCloy, Oppler, and Sager’s (1993) theory of performance, job knowledge is an important determinant of Soldier job performance. Tests of job knowledge are also more practical to administer than higher fidelity measures of job performance, such as hands-on tests. Additionally, JKTs tend to exhibit good test characteristics (e.g., internal consistency). Attrition occurs when Soldiers separate from the Army before completion of their contracted term of enlistment. Attrition is costly to the Army, greatly impacting the resources required for accessing and training a ready force.

Job Knowledge Tests (JKTs)

All Soldiers graduating from schoolhouses participating in this research and Soldiers in the in-unit data collections are administered a selected response test of general Soldier knowledge (i.e., Warrior Tasks and Battle Drills [WTBD]). To support classification research, IMT Soldiers in 11 MOS also take an MOS-specific JKT. The validation analyses in the present report use MOS-specific JKTs administered during IMT to Soldiers in the following MOS:

- Infantryman (11B/C/X + 18X)
- Combat Engineer (12B)
- Cannon Crewmember (13B)
- Field Artillery Tactical Data Systems Specialist (13D)
- Fire Support Specialist (13F)
- Cavalry Scout (19D)
- Armor Crewman (19K)
- Military Police (31B)
- Combat Medic Specialist (68W)
- Motor Transport Operator (88M)
- Wheeled Vehicle Mechanic (91B)

JKT development (Graves & Taylor, 2020) followed the process established and documented over the course of several ARI-sponsored criterion development efforts, including Project A (Campbell & Knapp, 2001) and Army Class (Moriarty, Campbell, Heffner, & Knapp, 2009). All JKTs were originally developed in two versions: one for administration toward the end of IMT and another for administration to Soldiers once they were in their respective units. In 2016, administration of the MOS-specific JKTs to Soldiers who were in their units was suspended to reduce overall assessment administration time.

The WTBD and Infantry (11B) JKTs were recently updated to account for changes in equipment, training, and doctrine. At the same time, the IMT and in-unit versions of the WTBD were combined into a single test for administration during IMT and in-unit. This new WTBD test was also updated for changes in equipment, training, and doctrine. The updated tests were first administered in 2018.

The JKTs include roughly 30 scored items. Items can be worth multiple points, depending on item type (e.g., multiple-choice, multiple-response). We compute a single, overall raw score for each JKT by summing the total number of points earned across all items and computing a percent correct score based on the maximum number of points that could be obtained on each test.

Table 3.4 shows coefficient alpha reliability estimates for the WTBD and MOS-Specific JKTs. The MOS-specific reliability estimates were reasonably high (mean $\alpha = .81$). The WTBD reliability estimates were similar (IMT $\alpha = .81$, in-unit $\alpha = .79$); this represents a considerable improvement over the reliability estimates prior to revision (IMT $\alpha = .68$, in-unit $\alpha = .56$). The

revised MOS-specific JKT for Infantry (11B) showed a decrease in reliability (from $\alpha = .80$ to $\alpha = .71$) but remains in the acceptable range.

Table 3.4. Job Knowledge Test (JKT) Reliability Estimates in the IMT and In-Unit Validation Samples

Domain/JKT	<i>N</i>	α
<i>IMT</i>		
<i>MOS-Specific</i>		
Infantryman (11B/C/X + 18X)	24,771	.71
Combat Engineer (12B)	3,846	.80
Cannon Crewmember (13B)	435	.82
Field Artillery Tactical Data Systems Specialist (13D)	804	.77
Fire Support Specialist (13F)	1,066	.83
Cavalry Scout (19D)	2,313	.84
Armor Crewman (19K)	2,584	.84
Military Police (31B)	14,671	.81
Combat Medic Specialist (68W)	14,340	.78
Motor Transport Operator (88M)	9,527	.80
Wheeled Vehicle Mechanic (91B)	3,055	.86
<i>WTBD (Army-Wide)</i>	30,413	.81
<i>In-Unit</i>		
<i>WTBD (Army-Wide)</i>	1,693	.79

Note. α = Coefficient Alpha. WTBD = Warrior Tasks and Battle Drills. Reliability estimates were computed using the Cycle 21 database.

Table 3.5 shows descriptive statistics for the IMT MOS-specific JKTs and for the IMT and in-unit WTBD JKTs. The IMT JKT scores averaged roughly 61% correct. In-unit WTBD JKT scores were similar, with a rough average of 58% correct. The correlation between the IMT MOS-specific JKTs taken together and the WTBD JKT was .50 indicating that they were related but provided unique information about Soldier performance.

Table 3.5. Job Knowledge Test (JKT) Descriptive Statistics in the IMT and In-Unit Validation Samples

Domain/JKT	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>	<i>r_{WTBD}</i>	<i>r_{AFQT}</i>
IMT							
<i>MOS-Specific</i>							
Infantryman (11B/C/X + 18X)	13,523	59.77	11.71	23.33	93.48	0.56	0.41
Combat Engineer (12B)	3,233	57.25	16.78	11.43	97.14	0.58	0.50
Cannon Crewmember (13B)	362	56.50	17.66	11.11	91.67	0.52	0.33
Field Artillery Tactical Data Systems Specialist (13D)	703	49.66	15.23	8.11	81.08	0.65	0.39
Fire Support Specialist (13F)	847	63.14	17.53	14.71	97.06	0.58	0.32
Cavalry Scout (19D)	1,981	48.22	13.84	10.34	77.59	0.57	0.29
Armor Crewman (19K)	1,974	65.99	14.23	18.64	96.61	0.60	0.25
Military Police (31B)	9,112	62.34	10.31	25.44	91.26	0.58	0.38
Human Resources Specialist (42A)	1,108	55.13	12.53	15.09	86.79	0.56	0.43
Combat Medic Specialist (68W)	7,504	71.83	8.98	23.47	92.86	0.44	0.27
Motor Transport Operator (88M)	5,234	62.03	10.15	25.71	88.89	0.58	0.41
Wheeled Vehicle Mechanic (91B)	1,890	56.30	12.84	23.40	90.72	0.52	0.29
All MOS Combined	49,139	60.78	14.13	0.00	97.14	0.50	0.35
<i>WTBD (Army-Wide)</i>	78,295	57.40	15.35	2.70	100.00		0.35
In-Unit							
<i>WTBD (Army-Wide)</i>	6,467	58.41	13.83	8.11	96.15		0.38

Note. *Ms*, *SDs*, *Min*, and *Max* reflect percent correct. IMT = Initial Military Training. WTBD = Warrior Tasks and Battle Drills. *r_{WTBD}* = correlation with WTBD JKT scores. *r_{AFQT}* = correlation with AFQT scores. Statistics for MOS with fewer than 100 cases are not reported. All correlations are statistically significant ($p < .05$, one-tailed).

Administrative Criteria

Administrative criteria used in VAST include attrition and number of training restarts.

Attrition is a broad category that encompasses involuntary and voluntary separations for a variety of reasons (e.g., family concerns, drug or alcohol use, performance, physical standards or weight, or conduct). The reason for separation was determined by the Soldiers' Separation Program Designator (SPD) code. Soldiers who left the Army for reasons outside of their or the Army's control (e.g., death or serious injury incurred while performing one's duties) were excluded from our analyses. Additionally, Soldiers with SPD codes indicating re-enlistment or completion of service are not considered as having separated from the Army for the purposes of our analyses. Separation data are reported for Regular Army Soldiers only because reliable data for the Reserve and National Guard components are not available.

A Soldier's time in service is calculated as the difference between the date from which their TAPDB record was extracted and their "ship" date as indicated in the Military Integrated Resource System (MIRS). We create separate monthly attrition status variables through 24 months within a Soldier's time in service. For example, if a Soldier has a total of 12 months in service, they will be indicated as 0 (in service) or 1 (separated) for each month from first through 12th and indicated as missing data (cannot be determined) for each month from 13th through 24th. Their overall attrition status is taken as their most recent non-missing attrition status. Table 3.6 summarizes basic descriptive statistics for attrition.

Soldiers' IMT completion status and whether they graduated from IMT with training recycles or restarts are extracted from ATRRS. IMT Restarts identify Soldiers who had to begin Basic Combat Training (BCT) or AIT again. Soldiers who had not had an opportunity to complete their training at the time data were extracted are excluded from analyses.

Table 3.6. Base Rates for Dichotomous Criteria in the Administrative Validation Sample

<i>Cumulative Regular Army Attrition</i>	<i>n^a</i>	<i>n_{Attrit}</i>	<i>%Attrit</i>
6-Month	401,020	47,205	11.77
12-Month	389,972	57,990	14.87
24-Month	356,997	82,508	23.11
<i>Restarted Initial Military Training (IMT)</i>	<i>n^b</i>	<i>n_{Restarted}</i>	<i>%Restarted</i>
IMT Restarts	704,166	38,871	5.52

Note. n^a = number of Soldiers with attrition data at the time data were extracted. n_{Attrit} = number of Soldiers who attrited at the specified months of service. $\%Attrit$ = percentage of Soldiers who attrited through the specified months of service $[(n_{Attrit}/n) \times 100]$. n^b = number of Soldiers with IMT data at the time data were extracted. $n_{Restarted}$ = number of Soldiers who restarted at least once during IMT. $\%Restarted$ = percentage of Soldiers who restarted at least once during IMT $[(n_{Restarted}/n) \times 100]$. Standardization excludes MOS with insufficient sample size ($n < 15$).

Army Life Questionnaire

The Army Life Questionnaire (ALQ) is used to collect self-reported administrative data and attitudes. The questionnaire asks Soldiers to report on their attitudes toward the Army and their MOS, indicate their intentions for their future with the Army, and provide information related to their Army service (e.g., awards, proficiency test scores). The ALQ versions administered to Soldiers during training and for Soldiers in their units of assignment are highly similar, but adjusted to reflect these different career stages.

The ALQ was originally developed under the Select21 project (Knapp et al., 2005), though it had its origins in Project A (Campbell & Knapp, 2001). Earlier forms of the ALQ (Van Iddekinge, Putka, & Sager, 2005) were modified slightly for use in the TOPS IOT&E, then additional modifications were made during the course of the research to broaden the scope of the criterion space covered by the ALQ, most recently in Cycle 17. Those modifications are briefly described

in this chapter; for a full summary of the changes to the ALQ see Nesbitt, Salmon, and Kirkendall (2020).

The ALQ was initially used to gather self-report administrative information in several areas, including two of particular interest in the present research: disciplinary incidents and Army Physical Fitness Test (APFT) scores. Disciplinary Incidents is scored as a dichotomous variable indicating if there is or is not a “yes” response to any associated items (e.g., Article 15, formally counseled, or placed on restriction for problem behavior). The APFT, which Soldiers took at approximate 6-month intervals, is a single continuous score. It is indexed by the ability to perform certain numbers of push-ups and sit-ups, and time taken to complete a 2-mile run, adjusted for age. The Army recently introduced the Army Combat Fitness Test (ACFT) to replace the APFT, but as the ACFT is not yet required, these scores are not currently included on the ALQ.

The ALQ also contains items rated on Likert-type scales to measure outcomes related to satisfaction and fit, constructs thought to be predictive of retention behavior. This portion of the ALQ was revised to accommodate new content without appreciably increasing total administration time and to maximize consistency across the IMT and in-unit ALQs (Nesbitt et al., 2020). The addition of new constructs was intended to expand the criterion domain to include a wider variety of Soldiers’ contextual performance and attitudes and investigate their relationship to TAPAS. Each of these measures underwent an iterative process, involving psychologists and Army subject matter experts (SMEs), to develop behavioral or attitudinal statements, evaluate their relevance and importance, and establish suitable psychometric properties. ALQ composite scores were revised to reflect the changes to the scales that comprise the composites. The correlations between the variables of interest and the pre- and post-revision composites, respectively, suggested the formulations were sufficiently similar to justify combining them for analysis purposes. The ALQ scales and composites are listed in Tables 3.1 and 3.2. For a full summary of the revision effort and description of new scales, see Nesbitt et al. (2020).

Table 3.7 shows the internal consistency reliability estimates for the IMT and in-unit ALQ scales and composites, all of which are at acceptable levels. Table 3.8 provides descriptive statistics for selected self-report administrative criteria, as well as the Likert-based scales and composites. Intercorrelations among the ALQ scales computed on the IMT and In-Unit Validation Samples are shown in Appendix C (Table C.1).

Table 3.7. Army Life Questionnaire (ALQ) Scale and Composite Reliability Estimates in the IMT and In-Unit Validation Samples

Variable	IMT			In-Unit		
	Alpha	Number Scales/Items	<i>n</i>	Alpha	Number Scales/Items	<i>n</i>
<i>Composites</i>						
Commitment and Fit ^a	0.81	3	121,893	0.83	3	17,558
Peer Leadership ^a	0.81	4	53,019	0.81	4	4,980
Retention Cognitions	0.86	2	121,448	0.87	2	4,978
<i>Likert Scales</i>						
Affective Commitment	0.79	3	121,893	0.80	3	17,558
Army Fit	0.77	4	121,893	0.80	4	17,558
Army Life Adjustment	0.78	4	121,893	-	-	-
Counterproductive Soldier Behavior	0.83	9	53,020	0.77	9	4,980
Deployment Satisfaction	-	-	-	0.80	4	1,765
MOS Fit	0.86	5	121,893	0.90	5	17,558
MOS Satisfaction	-	-	-	0.84	5	17,558
MTL - Affective	0.83	3	53,020	0.83	3	4,980
MTL - Noncalculative	0.77	3	53,020	0.77	3	4,980
MTL – Social-Normative	0.77	3	53,019	0.76	3	4,980
Organizational Citizenship Behavior/Leadership	0.86	8	53,020	0.86	8	4,980
Resilience	0.80	7	53,020	0.78	7	4,980
Work/Life Balance	0.87	5	52,971	0.86	5	4,980

Note. MTL = Motivation to Lead. Reliability estimates were computed using the Cycle 20 database.

^a Alphas reflect the average of the composite scale alphas.

Table 3.8 Army Life Questionnaire (ALQ) Descriptive Statistics in the IMT and In-Unit Validation Samples

Variable	IMT					In-Unit				
	<i>n</i>	Mean	<i>SD</i>	Min	Max	<i>n</i>	Mean	<i>SD</i>	Min	Max
<i>Composites</i>										
Commitment and Fit	92,550	3.78	0.65	1.00	5.00	8,482	3.31	0.78	1.00	5.00
Peer Leadership	48,665	3.31	0.44	1.00	5.00	3,350	3.25	0.48	1.00	5.00
Retention Cognitions	92,196	3.59	1.08	1.00	5.00	3,348	2.87	1.32	1.00	5.00
<i>Likert Scales</i>										
Affective Commitment	92,550	3.64	0.82	1.00	5.00	8,482	3.16	0.97	1.00	5.00
Army Career Intentions	92,280	3.55	1.13	1.00	5.00	8,457	2.77	1.32	1.00	5.00
Army Fit	92,550	3.91	0.70	1.00	5.00	8,482	3.48	0.93	1.00	5.00
Army Life Adjustment	92,550	3.80	0.88	1.00	5.00	—	—	—	—	—
Counterproductive Soldier Behavior	48,666	1.93	0.63	1.00	5.00	3,350	2.04	0.61	1.00	5.00
Deployment Satisfaction	—	—	—	—	—	898	3.38	0.97	1.00	5.00
MOS Fit	92,550	3.78	0.81	1.00	5.00	8,482	3.28	0.96	1.00	5.00
MOS Satisfaction	—	—	—	—	—	8,482	3.21	0.97	1.00	5.00
MTL – Affective	48,665	3.46	0.81	1.00	5.00	3,350	3.42	0.86	1.00	5.00
MTL - Noncalculative	48,666	2.28	0.85	1.00	5.00	3,350	2.42	0.90	1.00	5.00
MTL – Social-Normative	48,666	4.02	0.69	1.00	5.00	3,350	3.89	0.76	1.00	5.00
Organizational Citizenship Behavior/Leadership	48,666	3.48	0.67	1.00	5.00	3,350	3.27	0.74	1.00	5.00
Reenlistment Intentions	92,347	3.63	1.11	1.00	5.00	3,349	2.98	1.41	1.00	5.00
Resilience	48,666	4.01	0.60	1.00	5.00	3,350	3.87	0.66	1.00	5.00
Work/Life Balance	48,638	3.15	0.87	1.00	5.00	3,350	2.88	0.94	1.00	5.00
<i>Administrative Criteria</i>										
Disciplinary Incidents (Y/N)	91,771	0.23	0.42	91,771	1.00	8,545	0.26	0.44	8,545	1.00
APFT Score	87,543	250.95	28.46	87,543	300.00	8,122	251.00	28.83	8,122	300.00

Note. MTL = Motivation to Lead.

Performance Rating Criterion Measures

Performance ratings are desirable primarily because they can comprehensively cover technical, motivational, and non-cognitive elements of performance. The latter areas are particularly important in VAST, given that these performance elements are what the TAPAS was designed to predict. Performance ratings also provide an element of face validity that is appealing to policymakers. While it is true that performance ratings tend to have lower quality psychometric properties than objective (e.g., job knowledge tests) or self-report attitudinal measures, they still show sensible patterns of relationships with predictor measures.

Historically, VAST and TOPS IOT&E research has incorporated both supervisor and peer ratings of performance at IMT and in-unit; these sources have unique advantages and disadvantages. Supervisor ratings are logistically easier to collect than peer ratings because it is easier to identify suitable raters. However, it is difficult to obtain ratings from multiple

supervisor raters who are sufficiently familiar with Soldier performance and the time required for more senior personnel to complete the ratings can be prohibitive; the quality of supervisor ratings can suffer as a result. These circumstances led to the introduction of peer ratings fairly late in the TOPS IOT&E program and the decision to end the collection of IMT cadre ratings in VAST research as of Cycle 17. Currently, peer ratings are collected in both IMT and in-unit settings and supervisor ratings are collected in-unit only.

The remainder of this chapter first describes the Performance Rating Scales for peers (IMT and in-unit) and then supervisors. For a complete description of the development and revision of these measures see Knapp, Salmon, Kirkendall, and Burgoyne (2020).

Peer Performance Ratings

Peer performance ratings were introduced into this research in 2018. These efforts were primarily aimed at alleviating the burden associated with collection of ratings data from IMT cadre, but also to take advantage of the close daily interactions that peers have with each other during training and on the job.

IMT PRS-P

ARI researchers drafted the rating scale content and iteratively revised that content to refine the (a) rating dimensions, (b) question stem and example behaviors, and (c) response options. To align the IMT PRS-P with revisions to ALQ, they included scales to assess organizational citizenship behaviors (“Going Above and Beyond”), counterproductive behaviors (“Counterproductive Soldier Behaviors”), and resilience, which was eventually combined with Adjustment to the Army. The newly generated materials were subjected to retranslation by psychologists and review by Army SMEs.

All versions of the PRS include a “not observed” option on each rating scale and a single “familiarity” rating that allowed raters to indicate their opportunity to observe each Soldier being rated using a 4-point scale:

Indicate your opportunity to observe each Soldier you will be rating.

1. Not enough to judge the Soldier’s performance.
2. Enough to judge some aspects of the Soldier’s performance.
3. Enough to judge many aspects of the Soldier’s performance.
4. Enough to judge most aspects of the Soldier’s performance.

The final IMT PRS-P scales are presented in Table 3.9. The rating scales are all on a scale from one to five, but the anchors vary by dimension. Figure 3.1 presents two examples. Each scale also has a “Not Observed/Cannot Rate” response option.

Table 3.9. IMT PRS-P Rating Scale Dimensions and Associated Example Behaviors

Scale	Example Behaviors
Effort and Discipline	• Puts forth individual effort in study, practice, and other training activities • Shows respect in word and action towards superiors • Obeys instructions, rules, and regulations • Maintains self, uniforms, living areas, and barracks to Army standards
Working Effectively with other Soldiers	• Respects, assists, and cooperates with fellow Soldiers • Treats all Soldiers with courtesy and respect, regardless of gender, race, ethnicity, ability, or background differences • Communicates ideas and information clearly and directly
Physical Fitness	• Maintains and enhances physical readiness • Performs required physical tasks
MOS Qualification Knowledge and Skill	• Learns and demonstrates AIT/OSUT knowledge and skills required for MOS qualification • Effectively applies training, rules, and strategies to solve problems
Resilience and Adjustment	• Persists in carrying out difficult tasks, even in stressful circumstances • Remains calm during high-pressure situations • Bounces back after facing a difficult situation or failure • Adjusts well to life in the Army
Counterproductive Soldier Behaviors	• Takes shortcuts that may be harmful • Takes unauthorized breaks • Leaves a mess for someone else to clean up
Going Above and Beyond	• Notices when a fellow Soldier is falling behind and offers assistance • Offers to help fellow Soldiers struggling to learn training material and skills • Goes out of the way to give another Soldier encouragement or express appreciation • Spends free time learning about procedures, equipment, etc.
Overall Performance	Instructions: Considering your evaluation of your peers on the dimensions important to being a successful Soldier, please rate the overall performance of each Soldier.

A. Effort and Discipline				
Example behaviors:				
• Puts forth individual effort in study, practice, and other training activities • Shows respect in word and action towards superiors • Obeys instructions, rules, and regulations • Maintains self, uniforms, living areas, and barracks to Army standards				
1	2	3	4	5
Much less than other Soldiers in my unit	Somewhat less than other Soldiers in my unit	About as much as other Soldiers in my unit	Somewhat more than other Soldiers in my unit	Much more than other Soldiers in my unit
B. Counterproductive Soldier Behaviors				
Example behaviors:				
• Takes shortcuts that may be harmful • Takes unauthorized breaks • Leaves a mess for someone else to clean up				
1	2	3	4	5
Much less often than other Soldiers in my unit	Somewhat less often than other Soldiers in my unit	About as often as other Soldiers in my unit	Somewhat more often than other Soldiers in my unit	Much more often than other Soldiers in my unit

Figure 3.1. Sample IMT peer rater performance rating scales.

In-Unit PRS-P

ARI researchers also developed in-unit peer rating scales (Knapp et al., 2020). They used the same design and basic content. They adapted the example behaviors to the in-unit context and added two new scales: “Processing Information and Solving Problems” and “Emergent Leadership.” Similar to the IMT PRS-P development effort, the new materials were retranslated by psychologists and reviewed by Army SMEs.

The final in-unit PRS-P rating dimensions and example behaviors are presented in Table 3.10. The in-unit PRS-P includes the same familiarity question and “Not Observed/Cannot Rate” response option as the other PRS.

Table 3.10. In-Unit PRS-P Rating Scale Dimensions and Associated Example Behaviors

Scale	Example Behaviors
Effort and Discipline	• Puts forth individual effort in completing tasks and handling responsibilities • Shows respect in word and action towards superiors • Obeys instructions, rules, and regulations • Maintains self, uniforms, and work space to Army standards
Working Effectively with other Soldiers	• Respects, assists, and cooperates with fellow Soldiers • Treats all Soldiers with courtesy and respect, regardless of gender, race, ethnicity, ability, or background differences • Communicates ideas and information clearly and directly
Physical Fitness	• Maintains and enhances physical readiness • Performs required physical tasks
MOS Qualification Knowledge and Skill	• Demonstrates knowledge and skills required for MOS • Handles most MOS task-related problems effectively • Effectively applies training, previous experience, rules, and strategies to complete job tasks
Processing Information and Solving Problems	• Efficiently handles day-to-day information load • Develops original and creative approaches to dealing with problems • Can integrate information from multiple sources • Considers relevant costs and benefits of alternative solutions
Resilience and Adjustment	• Persists in carrying out difficult tasks, even in stressful circumstances • Remains calm during high-pressure situations • Bounces back after facing a difficult situation or failure • Adjusts well to life in the Army
Counterproductive Soldier Behaviors	• Takes shortcuts that may be harmful • Takes unauthorized breaks • Leaves a mess for someone else to clean up • Speaks poorly about the Army to others
Going Above and Beyond	• Volunteers for additional duty • Goes out of the way to give another Soldier encouragement or express appreciation • Demonstrates concern about the image of the unit
Emergent Leadership	• Notices when a fellow Soldier is falling behind • Keeps the unit focused on goals and objectives even when supervisors are not present • Gains the cooperation and trust of other Soldiers in the unit
<i>Overall Performance</i>	Instructions: Considering your evaluation of your peers on the dimensions important to being a successful Soldier, please rate the overall performance of each Soldier.

PRS-P Scoring and Interrater Reliability

A Soldier's PRS-P ratings were not included in the analyses if the rater (a) indicated he or she had little opportunity to observe this Soldier (familiarity rating = 1), (b) omitted more than 10%

of the assessment items, or (c) indicated that he or she had not observed the Soldier on more than 50% of the dimensions.

Table 3.11 presents the interrater reliability estimates for the IMT and in-unit PRS-P. The estimates are higher for the IMT ratings than the in-unit ratings at least in part because there were generally more peer raters in IMT. This finding may also reflect differential opportunity to observe across the two environments – peers going through intensive training together compared to peers performing their jobs in the field.

Overall, the interrater reliability estimates decreased slightly since the scales were first administered in Cycle 17. Of note when reviewing these results is the Counterproductive Soldier Behaviors scale. This scale is reverse scored and, although raters are alerted to the fact that the anchors are not the same across all dimensions, it is likely that this is a factor in the especially low interrater reliability estimates for that scale.

Table 3.11. PRS-P Interrater Reliability Estimates in the IMT and In-Unit Validation Samples

Measure	$G(q,k)$	k	n_{Ratees}	n_{Raters}
<i>IMT</i>				
Effort and Discipline	0.41	2.24	46,291	41,834
Working Effectively with Other Soldiers	0.37	2.25	46,303	41,835
Physical Fitness	0.51	2.24	46,281	41,802
MOS Qualification Knowledge & Skill	0.38	2.24	46,303	41,789
Resilience and Adjustment	0.37	2.24	46,264	41,784
Counterproductive Soldier Behaviors	0.17	2.24	46,224	41,670
Going Above and Beyond	0.32	2.24	46,269	41,803
Overall Performance Rating	0.46	2.25	46,242	41,680
<i>In-Unit</i>				
Effort and Discipline	0.29	1.74	3,113	2,832
Working Effectively with Other Soldiers	0.19	1.74	3,115	2,840
Physical Fitness	0.39	1.74	3,121	2,838
MOS Qualification Knowledge & Skill	0.28	1.73	3,112	2,829
Processing Information & Solving Problems	0.24	1.74	3,113	2,834
Resilience and Adjustment	0.22	1.74	3,113	2,832
Counterproductive Soldier Behaviors	0.10	1.73	3,113	2,819
Going Above and Beyond	0.19	1.74	3,117	2,837
Emergent Leadership	0.28	1.73	3,111	2,834
Overall Performance Rating	0.31	1.74	3,110	2,834

Note. k = harmonic mean number of raters per ratee. Reliability for the PRS-P scales reflect $G(q,k)$ estimates (Putka, Le, McCloy, & Diaz, 2008). PRS ratings from peers with a familiarity rating of 1 (“I have had little opportunity to observe this Soldier”) were excluded from analyses. Reliability estimates were computed using the Cycle 21 database.

Given the relatively unreliability of the individual PRS-P scales, we calculate a single composite score to use as a criterion measure.

PSR-P Descriptive Statistics

Tables 3.12 and 3.13 report basic descriptive statistics for the IMT and in-unit PRS-P variables. The means on all scales were well above the mid-point of the scale (2.5), though the full range of the 5-point scale was used by some raters. Scale intercorrelations calculated on the IMT and In-Unit Validation samples are provided in Appendix C (Table C.2). The internal consistency (alpha) estimate for the Overall Composite rating was .92 for both the IMT and in-unit PRS-P.

Table 3.12. Descriptive Statistics for the Peer Performance Rating Scales (PRS-P) in the IMT Validation Sample

Domain/PRS	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>
Effort and Discipline	43,254	3.75	0.95	1	5
Working Effectively with Other Soldiers	43,317	3.76	0.94	1	5
Physical Fitness	43,211	3.68	0.91	1	5
MOS Qualification Knowledge & Skill	43,218	3.76	0.85	1	5
Resilience and Adjustment	43,230	3.67	0.89	1	5
Counterproductive Soldier Behaviors	42,975	2.18	1.03	1	5
Going Above & Beyond	43,179	3.51	0.96	1	5
Overall Performance Rating	43,339	3.75	0.88	1	5
Overall Performance Composite	43,372	3.48	0.61	1	5

Note. Ratings on PRS-P range from 1 and 5. PRS ratings from peers with a familiarity rating of 1 (“I have had little opportunity to observe this Soldier”) were excluded from analyses.

Table 3.13. Descriptive Statistics for the Army-Wide Performance Rating Scales (PRS-P) in the In-Unit Validation Sample

Domain/PRS	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>
Effort and Discipline	2,741	3.87	0.92	1	5
Working Effectively with Other Soldiers	2,742	3.96	0.87	1	5
Physical Fitness	2,729	3.70	0.93	1	5
MOS Qualification Knowledge & Skill	2,711	3.83	0.90	1	5
Information Processing	2,712	3.71	0.86	1	5
Resilience and Adjustment	2,714	3.78	0.88	1	5
Counterproductive Soldier Behaviors	2,708	3.37	1.28	1	5
Going Above & Beyond	2,717	3.52	0.97	1	5
Emergent Leadership	2,715	3.64	0.99	1	5
Overall Performance Rating	2,748	3.89	0.86	1	5
Overall Performance Composite	2,755	3.71	0.71	1	5

Note. Ratings on PRS-P range from 1 and 5. PRS ratings from peers with a familiarity rating of 1 (“I have had little opportunity to observe this Soldier”) were excluded from analyses.

In-Unit PRS-S

The PRS-S, like the JKTs, was adapted in TOPS IOT&E from, or based on, previous research dating back to Project A (Moriarty et al., 2009). The Army-wide rating scales are suitable for rating all Soldiers. The rating dimensions (described in Table 3.14) were selected based on a comprehensive set of performance dimensions identified in the Select21 research program (Knapp et al., 2005); they retranslated both the dimension definitions and behavioral statements that appears on the Select21 PRS, giving special attention to scales that had shown exceptionally low interrater reliability in prior research. For more detailed information about the content and psychometric properties of the scales, see Knapp et al. (2020).

Table 3.14. In-Unit PRS-S Army-Wide Rating Scale Dimensions and Associated Definitions

Performing Core Warrior Tasks	Performs most Core Warrior Tasks (e.g., navigation, first aid, weaponry, maintenance) competently and safely
Performing MOS-Specific Tasks	Performs MOS-specific work assignments; keeps informed of MOS and assignment changes
Communicating with Others	Speaks clearly and concisely; conveys intended message verbally and in writing
Processing Information	Monitors, interprets, and organizes information
Solving Problems	Adapts to new problem situations; applies prior training, rules, and strategies correctly; weighs alternatives when making decisions; develops novel solutions to problems; completes tasks despite major changes
Exhibiting Effort	Completes work in a timely manner; puts extra effort into completing work; seeks challenging assignments; persists in carrying out difficult assignments and responsibilities even under adverse and stressful conditions
Exhibiting Personal Discipline	Exhibits selfless service orientation; exhibits integrity and discipline both on and off the battlefield; follows instructions, rules, and regulations
Contributing to the Team	Treats team members courteously and respectfully; provides help and assistance to others; contributes to achieving team goals
Exhibiting Fitness and Bearing	Meets Army standards for physical fitness, strength, and weight; displays military bearing; meets Army standards for AR 670-1
Adjusting to the Army	Soldier fits within the Army; demonstrates adjustment to Army life, shows potential for promotion
Following Safety Procedures	Follows safety procedures; recognizes and responds to possible dangerous or hazardous situations
Developing Own Skills	Stays up to date with his or her professional skills by seeking out additional education and training opportunities; commits to learning new things required by technology, mission, or situation
Managing Personal Matters	Manages personal life, including personal finances, family, and personal well-being
Emergent Leadership	Potential effectiveness as an NCO

The in-unit PRS-S employs the 7-point BARS format (see Figure 3.2) and includes the same familiarity question and “Not Observed/Cannot Rate” response option as the other PRS. We average ratings on the individual scales to form a single Overall Performance Composite score. Data cleaning procedures for the PRS-S data parallel that used for the PRS-P data.

A. Performing Core Warrior Tasks						
Performs most Core Warrior Tasks (e.g., navigation, first aid, weaponry, maintenance) competently and safely.						
1	2	3	4	5	6	7
• Is not able to perform most Core Warrior Tasks. • Requires constant supervision.		• Performs most Core Warrior Tasks competently. • Requires some supervision under difficult conditions.			• Performs almost all Core Warrior Tasks extremely effectively. • Requires little or no supervision, even under difficult conditions.	

Figure 3.2. Sample in-unit supervisor rater performance rating scale.

The mean PRS-S composite score in the In-Unit Validation Sample is 4.85 ($SD = 1.13$, $n = 6,265$). The majority of Soldiers in units are rated by only one supervisor, so we cannot compute interrater reliability estimates.

Cross-Instrument Correlations and Subgroup Differences

Correlations among the IMT and in-unit variables included in validation analyses (see Chapter 4) are presented in Appendix C (Table C.3). Generally, correlations among the variables and across the instruments show logical patterns. Scores on knowledge and attitude measures (particularly those assessed in IMT) tend to have modest negative correlations with disciplinary action and attrition. In the in-unit PRS, where both peer and supervisor ratings are available, the correlation between those measures is quite strong ($r = .35$). Performance ratings from both IMT and in-unit settings tend to relate positively with the self-reported attitudinal and behavioral measures, with the exception of Counterproductive Soldier Behavior, as we might expect.

Group differences with respect to gender and race/ethnicity can be found in Appendix C (Tables C.4 and C.5, respectively). With respect to gender, there were few meaningfully sized differences. The largest differences were that females scored lower than males on Counterproductive Soldier Behavior both IMT and in-unit ($-.24$ and $-.32$, respectively). Females scored somewhat lower on the WTBD JKT and the Knowledge and Skill Composite, more so in-unit ($-.28$ and $-.26$, respectively) than in IMT ($-.19$ and $-.15$). The female group also tended to attrit at higher rates (effect sizes ranging from $.16$ to $.27$).

With respect to race and ethnicity, each of the minority groups tended to score lower on the Knowledge and Skill composite score and WTBD JKT, with effect sizes ranging from a high of $-.66$ (in-unit Black-White Knowledge and Skill composite comparison) to a low of $-.23$ (IMT Asian-White WTBD JKT comparison). Non-white Soldiers tended to be less likely to attrit but these effect sizes were small, ranging from $.02$ to $.08$ for Black-White comparisons, $.16$ to $.29$ for Asian-White comparisons, and $.11$ to $.19$ for Hispanic-White comparisons.

Summary

In this chapter, we have described the criterion measures used as indicators of Soldier outcomes in VAST. These criterion measures included objective indicators, self-report outcomes, and performance ratings.

Tests of job knowledge and administrative data provided objective indices of important aspects of Soldier success. The JKTs include a series of tests specific to selected MOS as well as an Army-wide test of knowledge of warrior tasks and battle drills. Recent revisions to the Army-wide WTBD JKT resulted in a significant improvement in the reliability estimates.

Administrative data included attrition and the number of training restarts, which provided additional indices of Soldier performance. Recent work to update the creation of attrition variables yielded estimates that are more consistent with publicly reported figures, with little impact on the magnitude of the rates and relationships.

The ALQ provided a measure of self-reported attitudes and administrative data. Recent efforts have expanded the criterion domain of behaviors and attitudes captured by the ALQ to increase the instrument's relevance and utility for future validations. Both the new and shortened scales continue to show strong reliability.

As part of VAST we collected performance ratings data from in-unit supervisors as well as peers of both IMT and in-unit Soldiers. Low interrater reliability for the individual scale ratings prompted us to compute a single overall rating composite. Relatively high scale intercorrelations among those scale ratings resulted in overall performance composite scores which have high internal consistency. While we have evidence of consistency across raters, however, this is weaker than desired and has not improved over time.

CHAPTER 4: TAPAS CRITERION-RELATED VALIDITY EVIDENCE

Tetsuhiro Yamada and Michael G. Hughes (HumRRO)

This chapter describes our criterion-related validation of TAPAS to predict Soldiers' performance and retention through their first enlistment term. For these analyses, we focus on the Can-Do, Will-Do, and Adaptation TAPAS composites, and their relationships with various cognitive, motivational, and attrition criteria. Consistent with the Army's personnel goals, we examined performance and retention-related outcomes that, as a group, provide representative coverage of the criterion space (Campbell, Hanson, & Oppler, 2001; Campbell, McHenry, & Wise, 1990; Knapp & Tremble, 2007; Strickland, 2005). We used regression analysis to examine the incremental validity of the TAPAS to predict outcomes beyond AFQT.

We evaluated TAPAS using a common set of criterion measures. Collectively, the measures represent a range of important Soldier outcomes and reflect both cognitive and non-cognitive indices, including attitudinal measures (i.e., ALQ) and performance ratings. Refer to Tables 3.1 through 3.3 in Chapter 3 which summarize the IMT and in-unit criteria, including a brief description of each.

While the three TAPAS composite scores were each designed to predict different outcomes (e.g., Adaptation is intended to predict attrition), we examine results for the full set of criterion measures for all three TAPAS composite scores.

Bivariate Correlations

Uncorrected zero-order correlations among the AFQT, TAPAS facet and composite scores, and continuous criteria are presented in Appendix D (Table D.1 for IMT criteria and Table D.2 for in-unit criteria.) Table D.3 presents comparable results for attrition outcomes. Despite the range restriction that presumably attenuates the AFQT estimates, the largest observed correlations are between AFQT and job knowledge test scores ($r = .43$ for both IMT and in-unit). Reflecting their design, the TAPAS Can-Do composite has its strongest relationship with job knowledge scores, while the strongest correlations with Will-Do and Adaptation composites are associated with several motivational criteria (e.g., Peer Leadership, Resilience).

Incremental Validity Analyses

The purpose of the incremental validity analyses is to examine the potential of TAPAS composites to predict Soldier outcomes beyond AFQT. For each TAPAS composite and criterion of interest, we used hierarchical regression to examine the incremental validity of TAPAS after accounting for the effect of AFQT. This approach is generally consistent with previous evaluations of TAPAS and similar experimental non-cognitive predictors (e.g., Hughes, 2020; Ingerick et al., 2009; Knapp & Heffner, 2009; 2010; Trippe, Caramagno, Allen, & Ingerick, 2011).

To conduct the hierarchical regression analyses, we regressed the criterion onto Soldiers' AFQT scores in Step 1, followed by the given TAPAS composite scores in Step 2 for each regression model. By using the TAPAS composites in the analyses, these models aim to reflect potential operational use of the TAPAS. Results of these analyses are presented in both tabular and

graphic form. Additionally, we took a second approach in which we entered individual component facets of the focal TAPAS composite. This approach provides a more theoretical perspective of the relationships between TAPAS *facets* and key criteria. The results of this approach are presented graphically. While sample sizes for all IMT analyses are quite large, they are relatively smaller for analyses involving the Can-Do and Adaptation TAPAS composites compared to the Will-Do composite because TAPAS v7 and v8 did not include all the facets required to compute these scores using the same formula. Additionally, sample sizes for the in-unit analyses are much smaller than for the IMT analyses. Even the smallest sample size is larger than 2,500, however, which is sufficient to provide reasonably stable results.

For all hierarchical regression analyses, we used Ordinary Least Squares (OLS) regression to examine continuously scaled criteria, and logistic regression for dichotomous criteria (i.e., attrition, Disciplinary Incidents, IMT Restarts). We discuss the results of the OLS regression analyses with respect to *R* values when interpreting the validity of TAPAS. We refer primarily to odds ratios (ORs) when interpreting the results of the logistic regression analyses. We provide additional details concerning the specific indices presented for the logistic regression analyses with the results of the dichotomous outcomes later in the chapter. We examined attrition criteria for Regular Army Soldiers only.

IMT Incremental Validity Analyses

Table 4.1 and Figures 4.1 to 4.3 summarize the incremental validity results of the TAPAS composites for predicting IMT performance criteria over and above AFQT. Overall, the results suggest that both the Will-Do and Adaptation TAPAS composites can enhance the Army's ability to predict various important IMT outcomes.

For cognitive criteria, the TAPAS composites (Can-Do, Will-Do, and Adaptation) evidenced no notable increments over AFQT in predicting scores on the composite measure of Knowledge and Skill and the WTBD job knowledge test ($\Delta R_s = .00$). These results are not surprising given that AFQT is an established strong predictor of Can-Do outcomes.

For motivation-oriented performance criteria, the TAPAS Will-Do composite enhanced the prediction of the APFT score ($\Delta R = .19$), Resilience scale ($\Delta R = .17$), and Peer Leadership composite ($\Delta R = .19$). Although to a lesser extent, it also provided substantial incremental validity for Counterproductive Soldier Behavior ($\Delta R = .06$), Commitment and Fit ($\Delta R = .09$), and the Peer Performance Rating composite ($\Delta R = .08$). The relatively small incremental gain in predicting the performance ratings score may be due to the low interrater reliability of the measure.

The TAPAS Adaptation composite also demonstrated a substantial degree of incremental validity in the prediction of APFT score ($\Delta R = .11$) and to a lesser extent, Resilience ($\Delta R = .05$) and the Peer Performance Rating composite ($\Delta R = .06$). The Can-Do composite is not designed to predict Will Do criteria, and as such it did not demonstrate any meaningful incremental validity for the motivation-based outcomes.

Table 4.1. Incremental Validity Estimates for TAPAS Composites over AFQT for Predicting IMT Criteria

Criterion Measure/Model	<i>n</i>	AFQT <i>R</i>	AFQT + TAPAS <i>R</i>	ΔR
<i>Knowledge & Skill</i>				
Can-Do	40,422	.43	.43	.00
Will-Do	46,701	.43	.43	.00
Adaptation	43,532	.43	.43	.00
<i>WTBD JKT</i>				
Can-Do	69,853	.35	.35	.00
Will-Do	78,283	.35	.35	.00
Adaptation	74,039	.35	.35	.00
<i>APFT Score</i>				
Can-Do	78,694	.08	.08	.00
Will-Do	87,546	.08	.27	.19
Adaptation	83,141	.08	.18	.11
<i>Counterproductive Soldier Behavior</i>				
Can-Do	48,330	.01	.02	.01
Will-Do	48,646	.01	.07	.06
Adaptation	48,500	.01	.01	.00
<i>Resilience</i>				
Can-Do	48,330	.03	.04	.01
Will-Do	48,646	.03	.21	.17
Adaptation	48,500	.03	.09	.05
<i>Commitment and Fit</i>				
Can-Do	83,370	.04	.05	.01
Will-Do	92,537	.04	.13	.09
Adaptation	87,960	.04	.05	.01
<i>Retention Cognitions</i>				
Can-Do	83,052	.11	.11	.00
Will-Do	92,183	.11	.12	.01
Adaptation	87,629	.11	.11	.00
<i>Peer Leadership</i>				
Can-Do	48,329	.04	.06	.02
Will-Do	48,645	.04	.23	.19
Adaptation	48,499	.04	.04	.00
<i>PRS-P Overall Performance Composite</i>				
Can-Do	43,090	.07	.07	.00
Will-Do	43,355	.06	.14	.08
Adaptation	43,234	.07	.12	.06

Note. PRS = Performance Rating Scales, where -P denotes peer ratings. *R* = multiple correlations between the AFQT and selected TAPAS composites with the targeted criterion measure. ΔR = Increment in *R* from adding the selected TAPAS composite to the regression model [(AFQT + TAPAS) - AFQT Only]. Bolded values indicate $p < .05$.

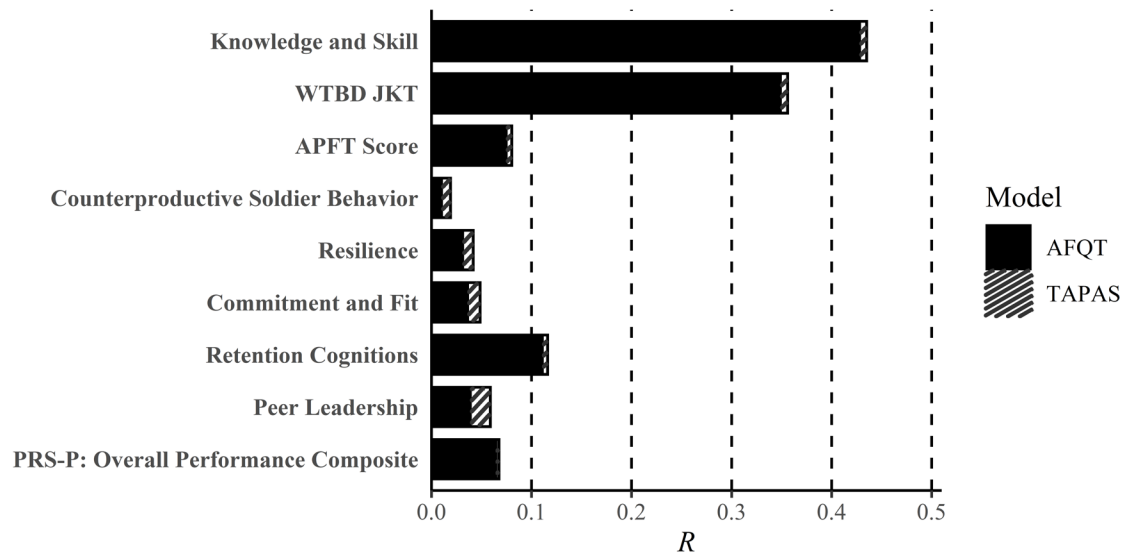


Figure 4.1. Increase in prediction of IMT criteria using TAPAS Can-Do composite.

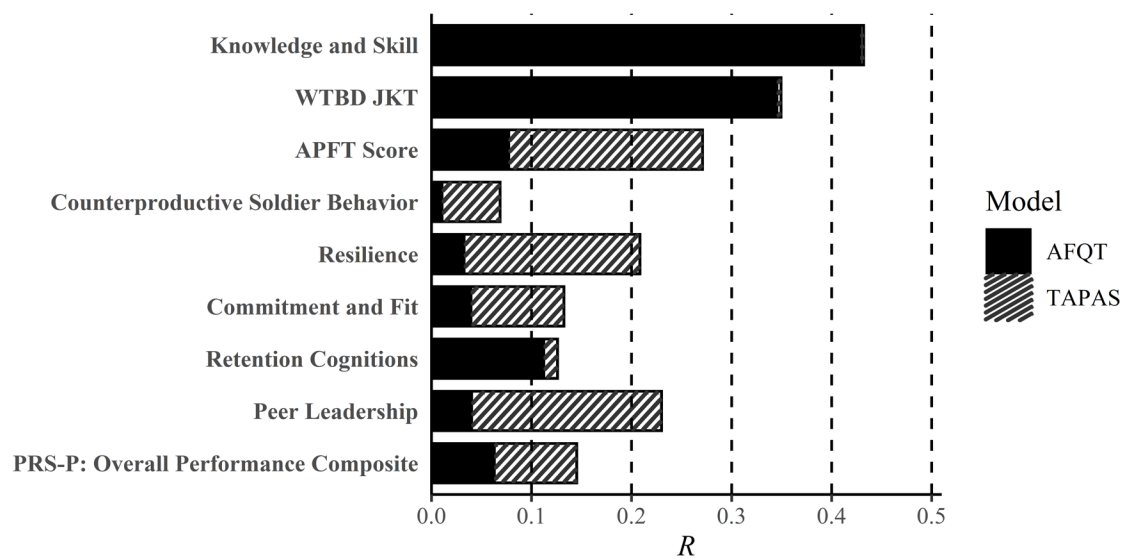


Figure 4.2. Increase in prediction of IMT criteria using TAPAS Will-Do composite.

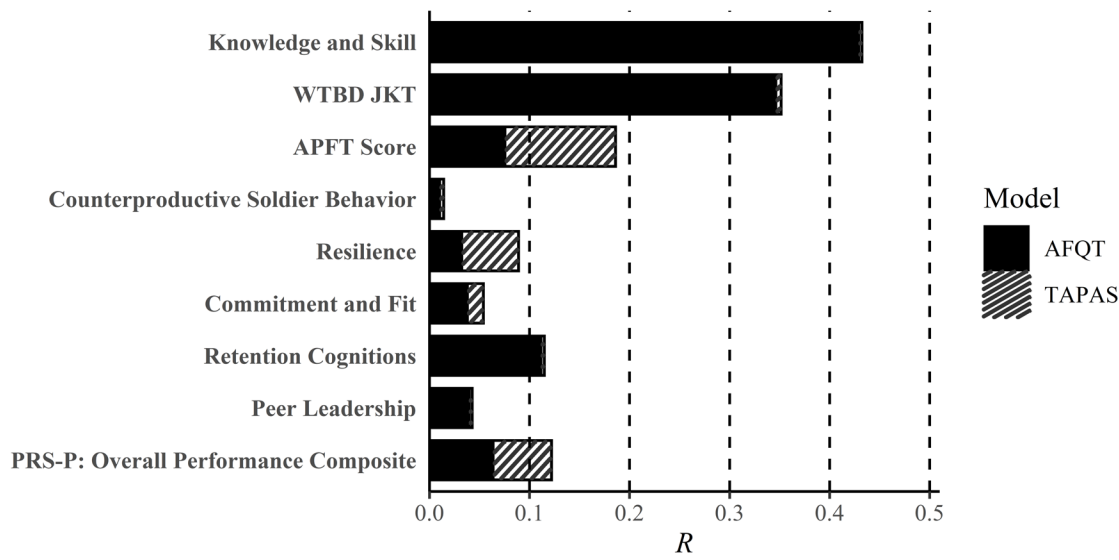
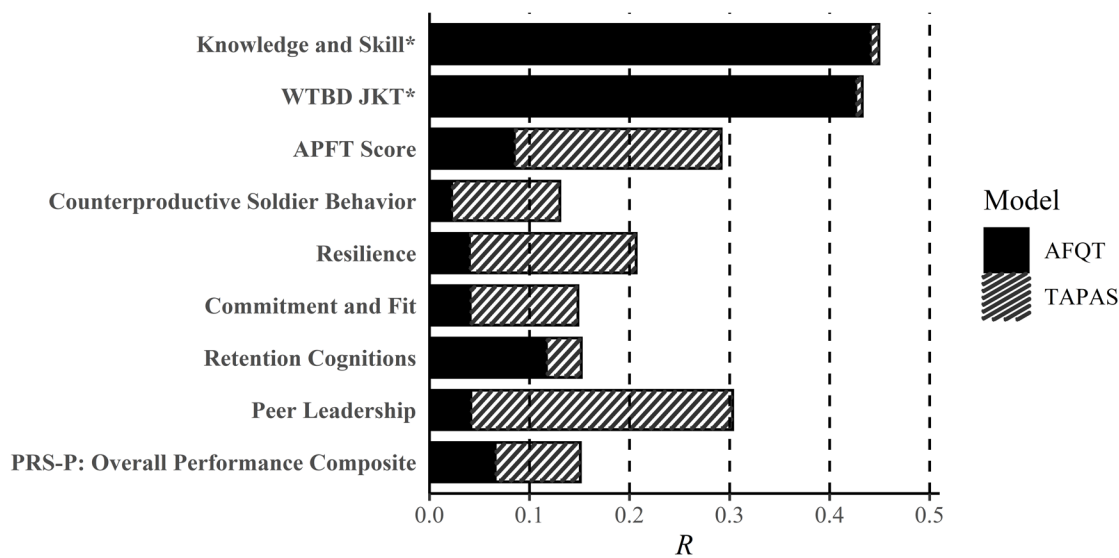


Figure 4.3. Increase in prediction of IMT criteria using TAPAS Adaptation composite.

Figure 4.4 provides results of analyses examining the incremental validity of criteria-specific TAPAS *facets* over and above AFQT for IMT Can Do and Will Do criteria. The results were similar to the TAPAS composite models discussed previously, in that Will-Do TAPAS facets provided larger incremental validity over AFQT than Can-Do facets. Can-Do TAPAS facets showed no notable increments over AFQT in predicting the Can Do criteria. Meanwhile, Will-Do TAPAS facets provided larger incremental validity, especially for Peer Leadership, Resilience, and APFT score ($\Delta R_s \geq .18$).



*Incremental validity of Can-Do facets is shown. For all other criteria, incremental validity of Will-Do facets is shown.

Figure 4.4. Increase in prediction of IMT criteria using relevant TAPAS facets.

In-Unit Incremental Validity Analyses

Table 4.2 and Figures 4.5 to 4.7 present incremental validity results for predicting in-unit performance. Similar to the results for the IMT performance criteria, these results also suggest that the TAPAS composites are useful predictors of in-unit performance outcomes. For multiple outcomes, the Will-Do and Adaptation TAPAS composites both exhibited enhanced prediction beyond AFQT alone.

Again, none of the TAPAS composites provided meaningful degree of incremental validity over AFQT in predicting job knowledge test scores.

For motivation-based criteria, the Will-Do composite provided incremental validity beyond AFQT in predicting APFT score ($\Delta R = .20$), Resilience ($\Delta R = .15$), Peer Leadership ($\Delta R = .15$), and to a lesser extent, the Peer Performance Rating composite ($\Delta R = .05$) scores. The TAPAS Adaptation composite demonstrated meaningful levels of incremental validity for predicting APFT score ($\Delta R = .13$) and Resilience ($\Delta R = .05$).

Table 4.2. Incremental Validity Estimates for TAPAS Composites over AFQT for Predicting In-Unit Criteria

Criterion Measure/Model	<i>n</i>	AFQT <i>R</i>	AFQT + TAPAS <i>R</i>	ΔR
<i>WTBD JKT</i>				
Can-Do	4,664	.37	.38	.01
Will-Do	6,467	.38	.38	.00
Adaptation	5,605	.38	.38	.00
<i>APFT Score</i>				
Can-Do	6,556	.01	.01	.00
Will-Do	8,634	.01	.22	.21
Adaptation	7,644	.00	.13	.13
<i>Counterproductive Soldier Behavior</i>				
Can-Do	3,508	.07	.08	.01
Will-Do	3,869	.08	.08	.01
Adaptation	3,701	.08	.08	.00
<i>Resilience</i>				
Can-Do	3,508	.02	.02	.01
Will-Do	3,869	.03	.18	.15
Adaptation	3,701	.03	.08	.05
<i>Commitment and Fit</i>				
Can-Do	6,852	.04	.04	.00
Will-Do	9,001	.03	.06	.02
Adaptation	7,980	.03	.04	.00
<i>Retention Cognitions</i>				
Can-Do	3,506	.09	.09	.00
Will-Do	3,867	.08	.08	.00
Adaptation	3,699	.09	.09	.00

Table 4.2. (Continued)

Criterion Measure/Model	<i>n</i>	AFQT <i>R</i>	AFQT + TAPAS <i>R</i>	ΔR
<i>Peer Leadership</i>				
Can-Do	3,508	.04	.05	.00
Will-Do	3,869	.05	.20	.15
Adaptation	3,701	.05	.05	.00
<i>PRS-P: Overall Performance Composite</i>				
Can-Do	2,596	.05	.05	.00
Will-Do	2,755	.05	.11	.05
Adaptation	2,679	.05	.07	.02
<i>PRS-S: Overall Performance Composite</i>				
Can-Do	4,734	.09	.09	.00
Will-Do	6,265	.09	.13	.03
Adaptation	5,542	.09	.10	.01

Note. PRS = Performance Rating Scales, where -S denotes supervisor ratings and -P denotes peer ratings. *R* = multiple correlations between the AFQT and selected TAPAS composites with the targeted criterion measure. ΔR = Increment in *R* from adding the selected TAPAS composite to the regression model [(AFQT + TAPAS) - AFQT Only]. Bolded values indicate $p < .05$.

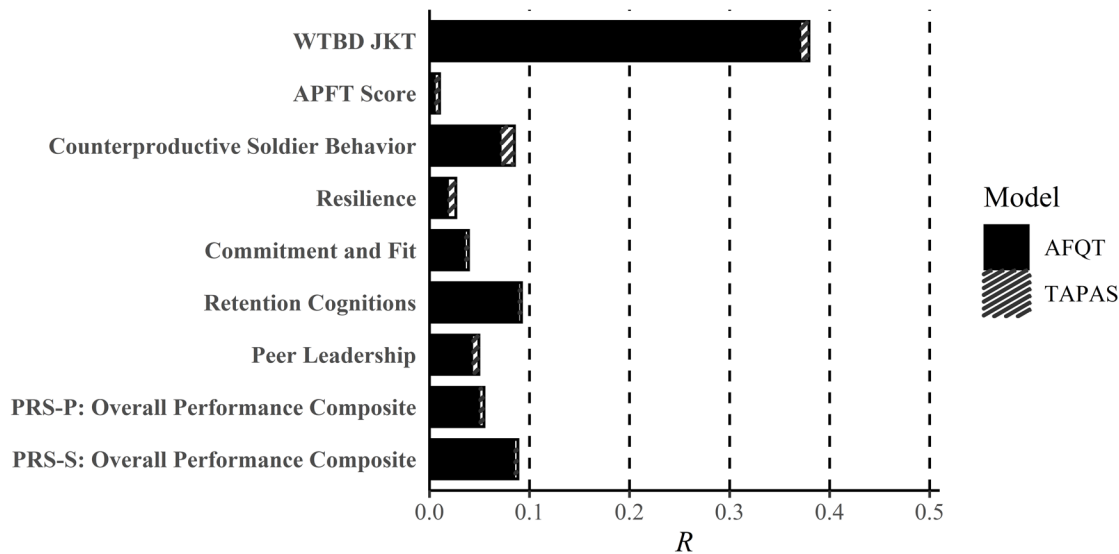


Figure 4.5. Increase in prediction of in-unit criteria using TAPAS Can-Do composite.

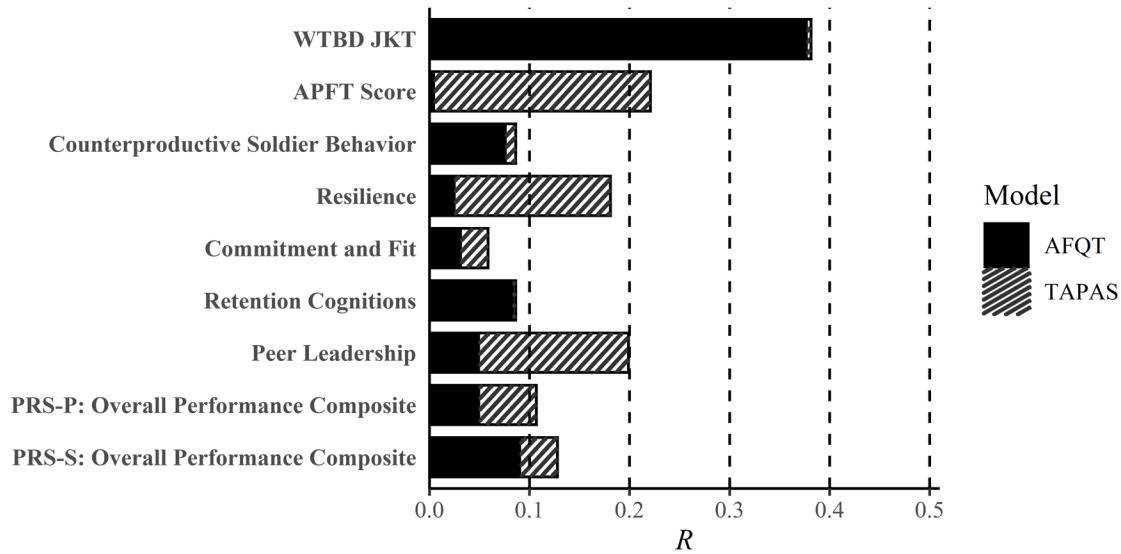


Figure 4.6. Increase in prediction of in-unit criteria using TAPAS Will-Do composite.

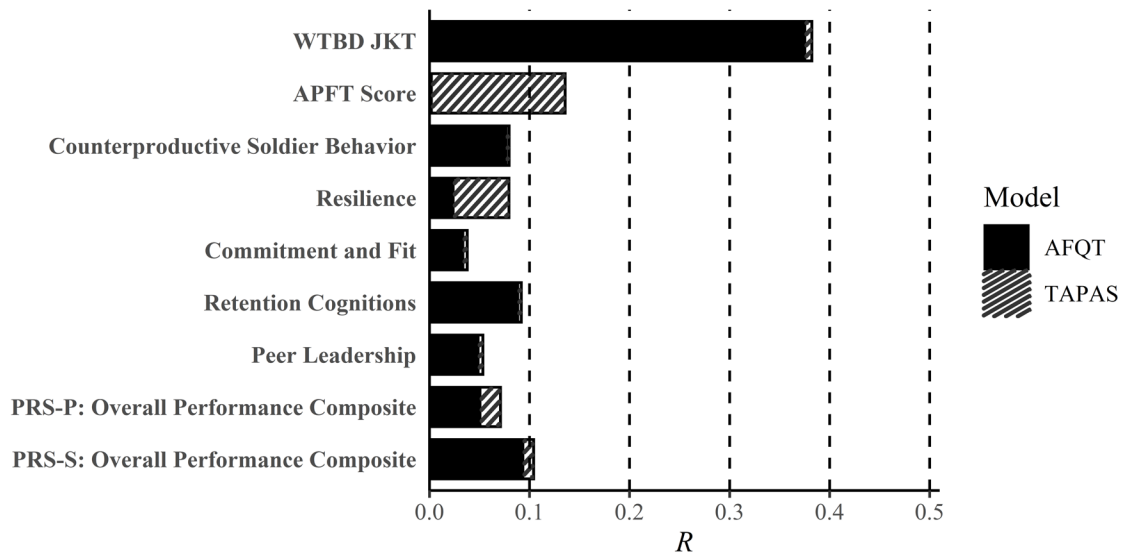
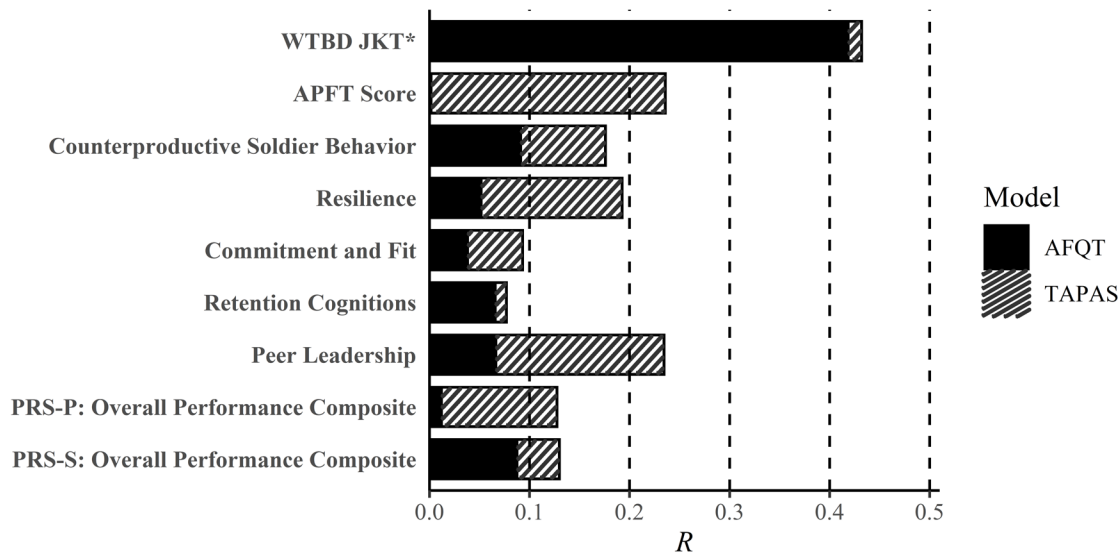


Figure 4.7. Increase in prediction of in-unit criteria using TAPAS Adaptation composite.

The results of the incremental validity analyses for individual TAPAS *facets* relevant to each criterion are shown in Figure 4.8. Similar to the results of the TAPAS composite models predicting both IMT and in-unit criteria, the Can-Do facets (i.e., those facets that are included in the Can-Do composite score) provided very little incremental validity over AFQT. On the other hand, Will-Do facets provided large gains over AFQT. Increases in R were largest for Peer Leadership, APFT, and Resilience ($\Delta R_s \geq .17$).



*Incremental validity of Can-Do facets is shown. For all other criteria, incremental validity of Will-Do facets is shown.

Figure 4.8. Increase in prediction of in-unit criteria using relevant TAPAS facets.

Predicting Dichotomous Outcomes

In addition to the OLS regression analyses of IMT and in-unit criteria, we conducted logistic regression analyses of the dichotomous outcomes, including Disciplinary Incidents, IMT Restarts, and attrition. For these models, we estimated odds ratios (ORs) for the predictors as well as the corresponding confidence intervals (CIs). Additionally, we computed point biserial correlations (r_{pb}) and conducted χ^2 tests of the change in model deviance (i.e., negative two log-likelihood; -2LL) between the AFQT-only and the AFQT + TAPAS composite models.

Odds ratios can be used to assess the likelihood (or odds) of the occurrence of a given outcome depending on change in a predictor. Specifically, for a given logistic regression model, a unique OR is estimated for each predictor, and represents the amount of change in the odds of the occurrence of the outcome that is associated with change in the given predictor. For the present analyses, the ORs represent the amount of change in the likelihood of each outcome that can be attributed to every 1.0 change in the predictor score. For ORs below 1.0, values closer to 0.0 indicate stronger negative relationships (i.e., lower likelihood of occurrence). Conversely, ORs above 1.0 indicate positive relationships (i.e., higher likelihood of occurrence). Although values of ORs cannot fall below 0.0, there is no upper limit for ORs (Cohen, Cohen, West, & Aiken, 2003). In addition, we computed 95% CIs for the ORs, which can be interpreted as an index of statistical significance for each. That is, a CI that contains 1.0 suggests that the relationship between the associated predictor and outcome is not statistically significant.

Point biserial correlations represent the correlation between a Soldier's predicted probability of exhibiting a selected behavior and his or her actual behavior (e.g., being involved in a disciplinary incident; Tabachnick & Fidell, 2013). As such, stronger point biserial correlations reflect stronger relationships between predicted and observed outcomes, and thus are indicative of better-fitting models. Model deviance (i.e., -2LL) also provides an index of model fit.

Moreover, the difference in deviances obtained from nested logistic regression models can be tested using likelihood ratio χ^2 tests to determine the statistical significance of change in model fit between models. In the present application, statistically significant likelihood ratio χ^2 tests of the change in deviances suggest that the inclusion of a given TAPAS composite to a regression model provides a significantly better prediction of the outcome than AFQT alone.

Results of the analyses examining Disciplinary Incidents (IMT and in-unit) and IMT Restarts are provided in Table 4.3. Both the Will-Do and Adaptation TAPAS composites enhanced the prediction of IMT Disciplinary Incidents and IMT Restarts beyond the AFQT-only models. Specifically, for both Disciplinary Incidents ($OR_{Will-Do} = .989$; $OR_{Adaptation} = .993$) and IMT Restarts ($OR_{Will-Do} = .994$; $OR_{Adaptation} = .994$), ORs associated with both composites were below 1.0, indicating that as scores on these composites increased, the likelihood of the outcome decreased. Similarly, for in-unit Disciplinary Incidents, both the Will-Do and Adaptation composites showed significant relationships ($OR_{Will-Do} = .993$; $OR_{Adaptation} = .995$) and resulted in better model fit over AFQT alone. The Can-Do composite did not predict either IMT or in-unit Disciplinary Incidents, as evidenced by CIs that include 1.0 for the associated ORs.

Table 4.3. Incremental Validity Estimates for TAPAS Composites over AFQT for Predicting IMT Restarts and Disciplinary Incidents

Criterion Measure/Model	<i>n</i>	$OR_{AFQT}(CI)$	$OR_{TAPAS}(CI)$	r_{pb}	$\Delta-2LL$
<i>IMT Restarts</i>					
Can-Do	575,503				
AFQT		.975 (.964-.986)		.01	
AFQT+TAPAS		.978 (.966-.990)	1.000 (.999-1.000)	.01	2.00
Will-Do	704,240				
AFQT		.978 (.969-.989)		.00	
AFQT+TAPAS		.988 (.978-.998)	.994 (.993-.994)	.03	601.15
Adaptation	639,940				
AFQT		.978 (.967-.988)		.01	
AFQT+TAPAS		.994 (.983-1.005)	.994 (.994-.995)	.03	467.53
<i>IMT Disciplinary Incidents</i>					
Can-Do	82,592				
AFQT		.988 (.972-1.004)		.01	
AFQT+TAPAS		.994 (.976-1.011)	.999 (.998-1.000)	.01	3.48
Will-Do	91,758				
AFQT		.985 (.970-1.000)		.01	
AFQT+TAPAS		.999 (.984-1.015)	.989 (.988-.990)	.09	784.58

Table 4.3. (Continued)

Criterion Measure/Model	<i>n</i>	OR _{AFQT} (CI)	OR _{TAPAS} (CI)	<i>r_{pb}</i>	Δ-2LL
<i>IMT Disciplinary Incidents (Continued)</i>					
Adaptation	87,181				
AFQT		.986 (.971-1.002)		.01	
AFQT+TAPAS		1.006 (.990-1.023)	.993 (.992-.994)	.06	304.73
<i>In-Unit Disciplinary Incidents</i>					
Can-Do	6,981				
AFQT		.962 (.909-1.018)		.02	
AFQT+TAPAS		.958 (.901-1.018)	1.001 (.998-1.004)	.02	0.15
Will-Do	9,163				
AFQT		.945 (.900-.992)		.02	
AFQT+TAPAS		.946 (.901-.993)	.993 (.991-.996)	.06	30.07
Adaptation	8,125				
AFQT		.947 (.898-.997)		.02	
AFQT+TAPAS		.959 (.909-1.010)	.995 (.992-.997)	.05	16.23

Note. OR = odds ratio for each predictor. CI = 95% confidence interval of the odds ratio. *r_{pb}* = point biserial correlation between the observed outcome and predicted probability. Δ-2LL = change in negative two log-likelihood (deviance) from adding the selected TAPAS composite score to the AFQT-only logistic regression model. Odds ratios equal to 1.0 (or confidence intervals of the odds ratio that include 1.0) indicate no relationship between the predictor and criterion. Odds ratios less than 1.0 indicate a negative relationship between the predictor and criterion. Odds ratios greater than 1.0 indicate a positive relationship between the predictor and criterion. For Δ-2LL, bolded indicate significant change in model fit based on a Likelihood Ratio χ^2 test, $p < .05$.

Figures 4.9 to 4.11 display the accuracy of the logistic regression models in predicting Disciplinary Incidents and IMT Restarts. For each outcome, results are presented for three models which include the following predictors: (a) AFQT, (b) TAPAS composite, and (c) AFQT + TAPAS composite. Specifically, the percent accurate values represent the area under the receiver operating characteristic (ROC) curve, which reflects the ability of the given model to correctly discriminate between a case and a noncase (e.g., attriter vs. stayer). The area under the curve (AUC) can range from 0 to 1.0, corresponding to predictions that are correct 0% and 100% of the time, respectively, with .5, or 50%, being chance. For example, the closer AUC is to 1.0, the more likely it is that Soldiers will be correctly predicted to have at least one IMT Disciplinary Incident versus no IMT Disciplinary Incidents.

For all three criteria, the probability of discriminating between Soldiers with and without incidence of the outcome of interest increased when either the Will-Do or Adaptation composite was added to the model. At the highest, such combined models demonstrated a prediction accuracy of 56.2% for IMT Disciplinary Incidents. Additionally, the difference in prediction accuracy between the Will-Do or Adaptation TAPAS composite-only and AFQT + TAPAS composite models was small,

suggesting that AFQT adds little value to the discrimination of these dichotomous criteria. However, results were different regarding the Can-Do composite. For both Disciplinary Incidents, use of the Can-Do composite in addition to AFQT resulted in very little change in predictive accuracy.

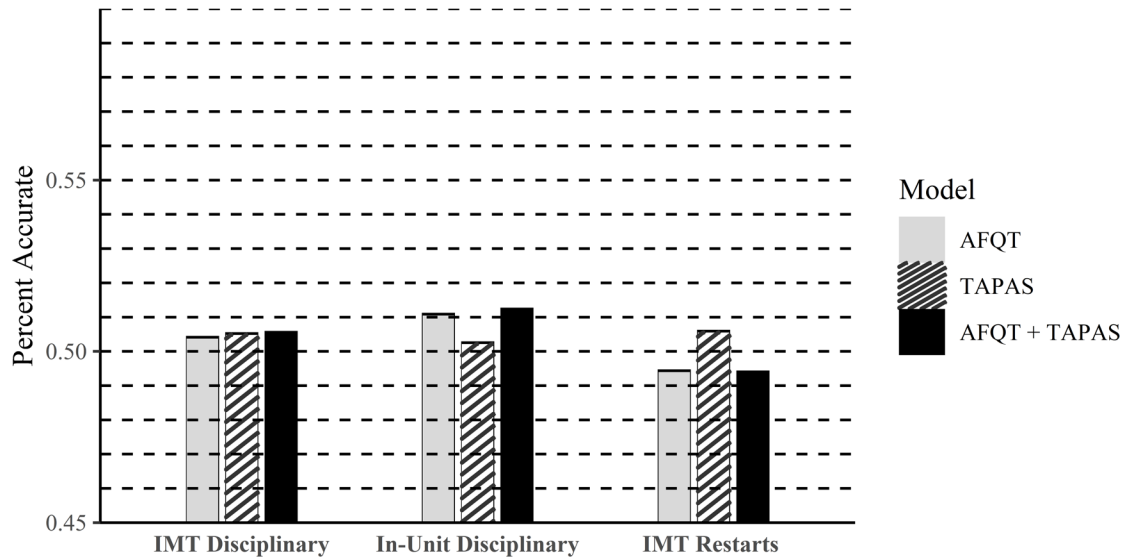


Figure 4.9. Predictive accuracy of AFQT and TAPAS Can-Do composite in the discrimination of both IMT and in-unit Disciplinary Incidents and IMT Restarts.

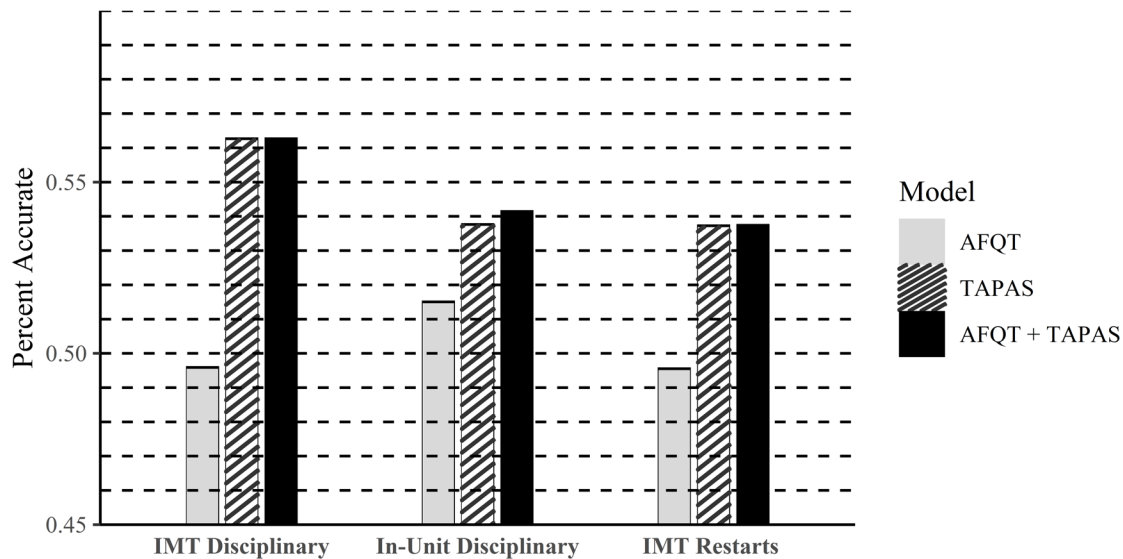


Figure 4.10. Predictive accuracy of AFQT and TAPAS Will-Do composite in the discrimination of both IMT and in-unit Disciplinary Incidents and IMT Restarts.

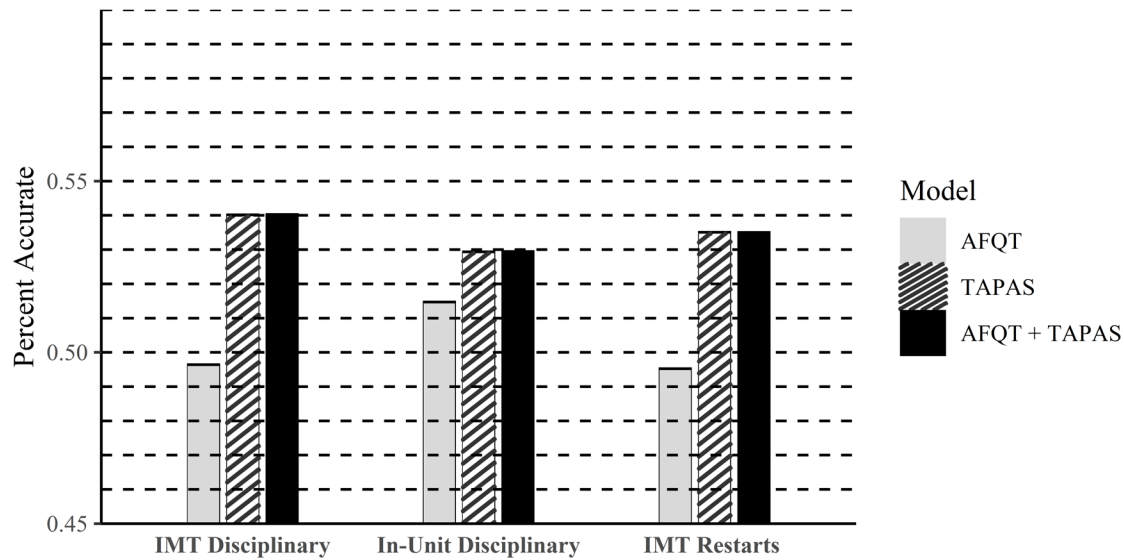


Figure 4.11. Predictive accuracy of AFQT and TAPAS Adaptation composite in the discrimination of both IMT and in-unit Disciplinary Incidents and IMT Restarts.

Figure 4.12 presents the accuracy of the logistic regression models in predicting Disciplinary Incidents and IMT Restarts when individual Will-Do TAPAS *facets* are used in place of the Will-Do composite. The results are similar to when the Will-Do composite is used such that the Will-Do facets alone produce much higher accuracy in predicting these criteria compared to AFQT alone, and combining the TAPAS facets and AFQT together do not lead to much improvement compared to using the TAPAS facets alone.

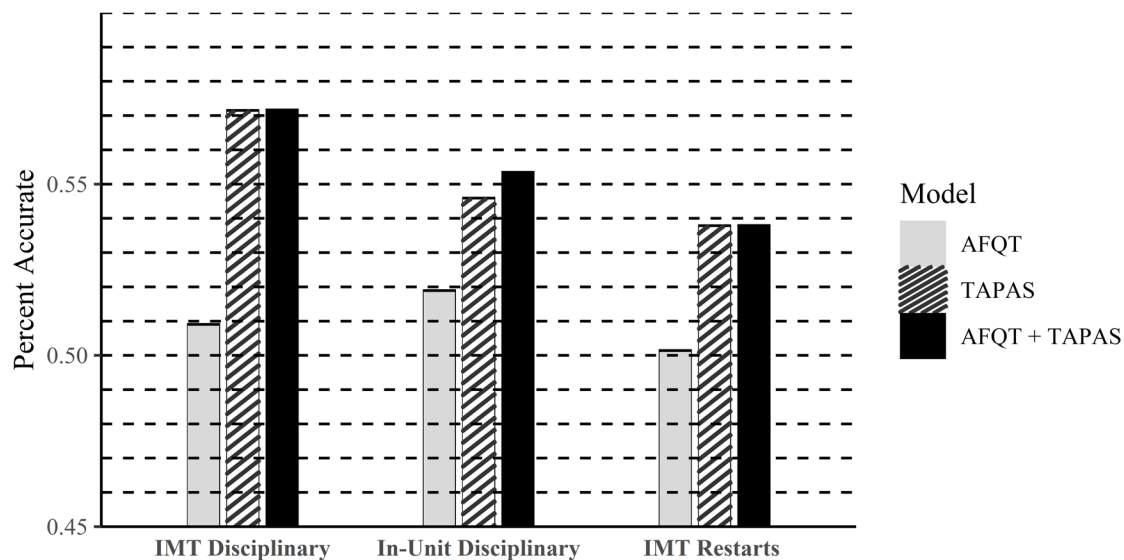


Figure 4.12. Predictive accuracy of AFQT and TAPAS Will-Do facets in the discrimination of both IMT and in-unit Disciplinary Incidents and IMT Restarts.

Table 4.4 presents the results of the logistic regression analyses examining attrition through 6, 12, and 24 months of service for Regular Army Soldiers. The TAPAS Will-Do ($OR_{Will-Do} = .988, .989, .991, \text{ and } .992$, respectively) and Adaptation ($OR_{Adaptation} = .989, .989, .990, \text{ and } .991$, respectively) composites were negatively related to attrition at all four time points, and their respective inclusion in the models resulted in significantly better fit over AFQT alone. Conversely, the TAPAS Can-Do composite was positively related to attrition at each of the time points for Regular Army Soldiers ($OR_{Can-Do} > 1.00$). That being said, the strength of the positive relation was weak, with point estimates being 1.001 and the upper bounds of these ORs were as high as 1.002.

Table 4.4. Incremental Validity Estimates for TAPAS Composites over AFQT for Predicting Cumulative Attrition through 24 Months of Service (Regular Army Only)

Criterion Measure/Model	<i>n</i>	$OR_{AFQT}(CI)$	$OR_{TAPAS}(CI)$	r_{pb}	$\Delta-2LL$
<i>6-month Attrition</i>					
Can-Do	334,793				
AFQT		.840 (.830-.849)		.05	
AFQT+TAPAS		.831 (.821-.841)	1.001 (1.001-1.002)	.05	22.83
Will-Do	401,044				
AFQT		.842 (.834-.850)		.05	
AFQT+TAPAS		.851 (.843-.860)	.988 (.988-.989)	.09	2262.04
Adaptation	367,866				
AFQT		.840 (.831-.848)		.05	
AFQT+TAPAS		.865 (.856-.874)	.989 (.989-.990)	.09	1814.50
<i>12-month Attrition</i>					
Can-Do	323,944				
AFQT		.870 (.861-.879)		.05	
AFQT+TAPAS		.862 (.853-.871)	1.001 (1.001-1.002)	.05	20.78
Will-Do	389,998				
AFQT		.873 (.865-.881)		.05	
AFQT+TAPAS		.881 (.873-.890)	.989 (.988-.989)	.09	2555.95
Adaptation	356,932				
AFQT		.871 (.863-.879)		.05	
AFQT+TAPAS		.897 (.889-.906)	.989 (.988-.989)	.09	2247.34

Table 4.4. (Continued).

Criterion Measure/Model	<i>n</i>	OR _{AFQT} (CI)	OR _{TAPAS} (CI)	<i>r_{pb}</i>	Δ-2LL
<i>24-month Attrition</i>					
Can-Do	291,230				
AFQT		.843 (.836-.851)		.07	
AFQT+TAPAS		.835 (.827-.843)	1.001 (1.001-1.002)	.07	30.24
Will-Do	357,022				
AFQT		.848 (.841-.855)		.07	
AFQT+TAPAS		.855 (.848-.862)	.991 (.991-.991)	.10	2054.54
Adaptation	324,077				
AFQT		.845 (.838-.853)		.07	
AFQT+TAPAS		.868 (.860-.875)	.990 (.990-.991)	.11	2238.84

Note. OR = odds ratio for each predictor. CI = 95% confidence interval of the odds ratio. *r_{pb}* = point biserial correlation between the observed outcome and predicted probability. Δ-2LL = change in negative two log-likelihood (deviance) from adding the selected TAPAS composite score to the AFQT-only logistic regression model. Odds ratios equal to 1.0 (or confidence intervals of the odds ratio that include 1.0) indicate no relationship between the predictor and criterion. Odds ratios less than 1.0 indicate a negative relationship between the predictor and criterion. Odds ratios greater than 1.0 indicate a positive relationship between the predictor and criterion. For Δ-2LL, bolded indicate significant change in model fit based on a Likelihood Ratio χ^2 test, $p < .05$.

Figures 4.13 to 4.15 present the accuracy of logistic regression models in predicting attrition criteria. As was the case with other criterion categories, the results were different between Can-Do and the other TAPAS composites. That is, the use of the TAPAS Will-Do or Adaptation composite in the logistic regression model resulted in higher prediction accuracy compared to the use of AFQT alone. Moreover, highest prediction accuracy was achieved when AFQT and either the Will-Do or Adaptation composite were concurrently used as predictors. Meanwhile, for every attrition criterion, the model with TAPAS Can-Do composite as the sole predictor had lower prediction accuracy than the AFQT-only model, and combination of AFQT and the Can-Do composite did not lead to substantial improvement in prediction accuracy compared to AFQT alone.

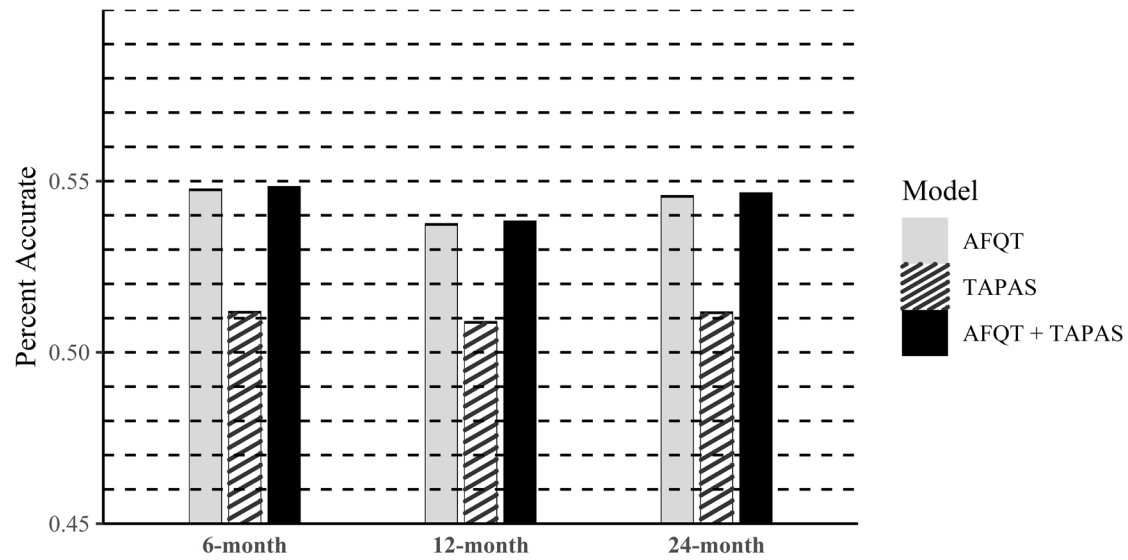


Figure 4.13. Predictive accuracy of AFQT and TAPAS Can-Do composite in the discrimination of Cumulative Attrition.

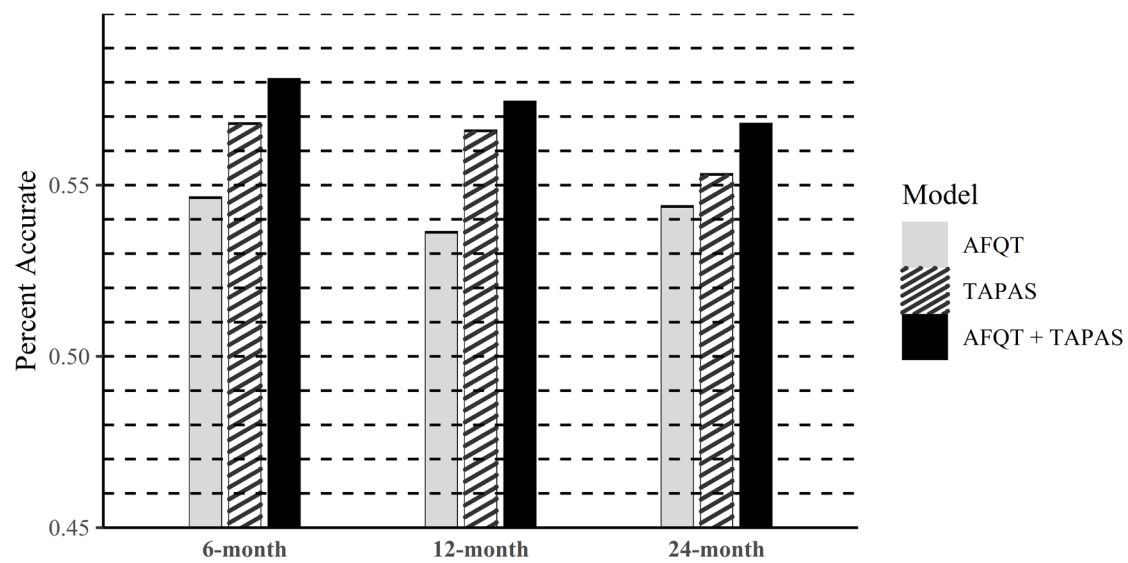


Figure 4.14. Predictive accuracy of AFQT and TAPAS Will-Do composite in the discrimination of Cumulative Attrition.

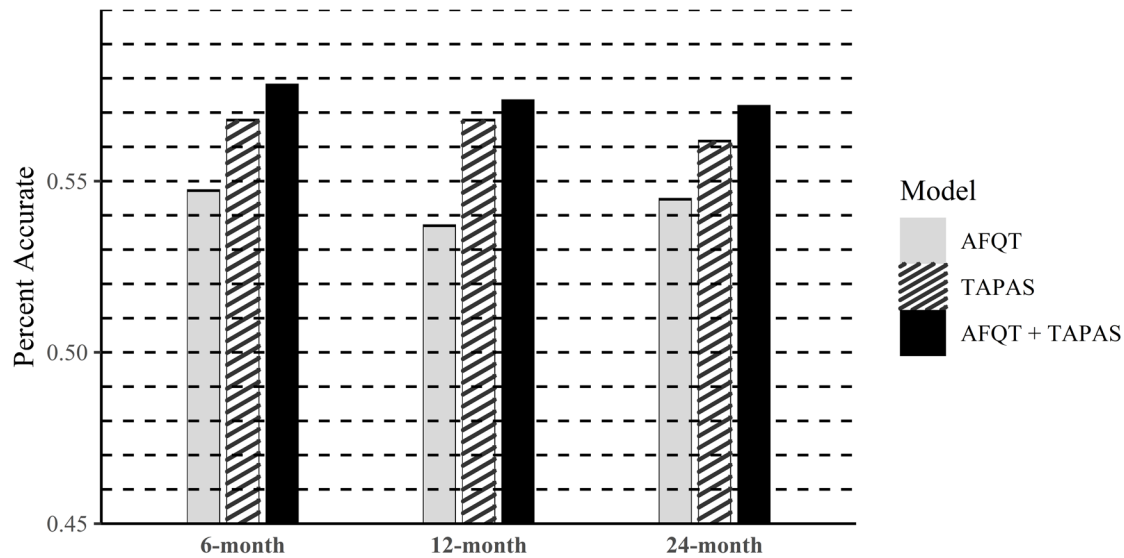


Figure 4.15. Predictive accuracy of AFQT and TAPAS Adaptation composite in the discrimination of Cumulative Attrition.

Figures 4.16 displays the accuracy of the logistic regression models in predicting attrition criteria when individual TAPAS Adaptation *facets* are included in the model in place of the Adaptation composite. The pattern of results largely mirrors what is presented in Figure 4.15 in that the Adaptation facets provide increased prediction accuracy over AFQT. However, the increase in accuracy is greater than that achieved by the composite. Most notably, the use of Adaptation facets in addition to AFQT results in an accuracy increase of 5.4% for predicting 12-month attrition.

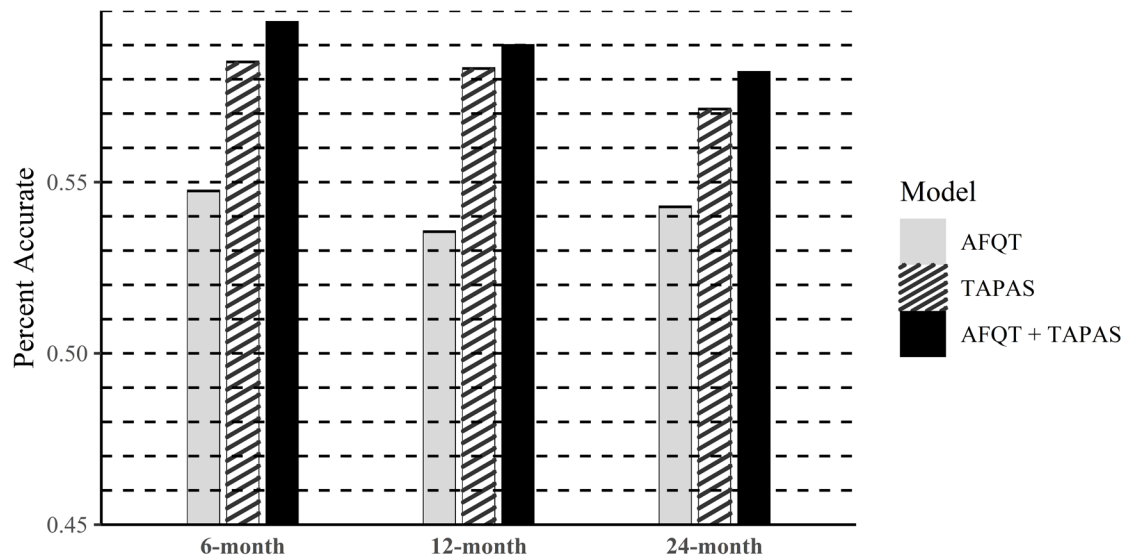


Figure 4.16. Predictive accuracy of AFQT and TAPAS Adaptation facets in the discrimination of Cumulative Attrition.

Summary

This chapter summarizes results from the 21st cycle of the evaluation of criterion-related validity of TAPAS. Results largely replicated those of prior analysis cycles such that Will-Do and Adaptation TAPAS composites demonstrated incremental validity over AFQT in predicting important first-term Soldier outcomes. For both IMT and in-unit outcomes, the Will-Do TAPAS composite demonstrated the strongest relationships with APFT score, Peer Leadership, and Resilience. More importantly, incremental validity analyses showed that both the Will-Do composite and its component facets explained additional variance in focal criteria above and beyond AFQT. Regarding dichotomous outcomes, addition of the Will-Do composite led to stronger prediction of IMT and in-unit Disciplinary Incidents and IMT Restarts compared to AFQT alone. The Adaptation composite also contributed to the prediction of attrition beyond AFQT alone. Given the well-established strength of AFQT to predict cognitive outcomes, it is unsurprising that TAPAS Can-Do composite and its component facets did not explain additional variance in criteria beyond AFQT.

CHAPTER 5: EXAMINATION OF SELECTED TEMPERAMENT CONSTRUCTS

Stephen A. Robertson and Samuel J. Posnock (HumRRO)

Introduction

Background

In line with HumRRO and ARI goals of conducting and supporting research efforts not only for operational use in the U.S. Army but also making contributions to the broader scientific literature, we describe a research effort using VAST data to examine research questions pertaining to the relationships between personality and performance. The VAST database is well positioned to answer these questions given that a large amount of data across many occupations is available. Additionally, analyses using TAPAS can shed light on recent questions regarding forced-choice personality measures due to its use of the Multi-Unidimensional Pairwise Preference (MUPP) Model, which is able to retrieve normative information from a format that often results in ipsative data when using other scoring models (Zhang et al., 2020).

Research Goal

The goal of the present research was to construct personality scores using TAPAS data that would facilitate comparison of personality-performance relations in a military setting with findings drawn largely from the civilian literature.

Personality Measurement

Personality assessments provide a promising technique to improve selection and job assignment in organizations. The development of the five-factor model introduced a framework to assess temperament systematically, driving advances in personality research (McCrae & Costa, 1987; Saucier & Goldberg, 1996). Scores on trait scales corresponding to each of the five factors have been shown to predict outcomes of interest to organizations, such as job performance, turnover, and counterproductive work behavior (CWB; Rothstein & Goffin, 2006). The dimensional structure has been widely replicated, and the predictive validity of the five-factor model persists across cultures and contexts (Barrick & Mount, 1991; Soto, 2019; Van Aarde, Meiring, & Wiernik, 2017).

The *Big Five* model emerged from a research context and, although it has found widespread adoption in industry, concerns persist over the validity of personality scores generated under high-stakes conditions. Conventional personality assessments use Likert-type response scales which pair a statement or adjective with polytomous or dichotomous response options. Under a classical test theory (CTT) paradigm, summing item scores yields a normative scale score for each broad domain (Brown & Maydeu-Olivares, 2011; 2013). Single-statement item formats may invite socially desirable responses, however, especially under competitive conditions such as high-stakes selection (Roulin, Krings, & Binggeli, 2016).

An alternative to Likert-type items is the forced-choice format in which respondents choose which statement or adjective, presented as multiple statements or adjectives, better reflects them. This approach may reduce motivated responding when the link to selection criteria is ambiguous or both attributes are socially desirable (Lee, Joo, & Lee, 2019; Salgado & Tauriz, 2014). Applying CTT scoring to forced-choice items results in ipsative scores or rank ordering of each personality subscale in the assessment. Such scale scores are useful for within-person comparisons but do not provide meaningful information for between-person comparisons (Lee et al., 2019).

TAPAS was developed to address the response distortion concerns associated with Likert-type scales and the ipsative scoring associated with forced-choice items (Drasgow et al., 2012). TAPAS applies an Item Response Theory (IRT) paradigm to forced-choice items by using a MUPP model (Drasgow et al., 2012; Lee et al., 2019). Statements within each item pair are matched on social desirability and extremity to minimize response distortion. The TAPAS assesses scores on narrow facets which can improve predictive validity beyond domain scores across specific job tasks (Judge, Rodell, Klinger, Simon, & Crawford, 2013).

Personality-Performance Relations

We had hoped to use TAPAS facets to construct scores on the Big Five personality variables. Given the composition of the various TAPAS versions on which we have data, however, we could construct reasonable variations of only two – Conscientiousness and Emotional Stability. Therefore, the following brief literature review focuses particularly on relations between these two aspects of personality and job performance.

Conscientiousness has been shown to predict work performance across multiple contexts. Barrick, Mount, and Judge. (2001) found a second-order meta-analytic relationship of $\rho = .27$ overall, with estimated validities for different performance elements ranging from $\rho = .31$ (supervisor ratings) to $\rho = .23$ (objective performance). Notably, Conscientiousness shows consistent results across occupations, with validities ranging from $\rho = .23$ for semi-skilled occupations to $\rho = .26$ for police officers (Barrick et al., 2001). Thus, we expect that Soldiers who tend to be ambitious, responsible, and orderly will perform more highly than those who score lower on these tendencies.

Emotional Stability has been shown to predict work performance generally, although findings are less consistent than for Conscientiousness. Barrick et al.'s (2001) second-order meta-analysis found a small, positive relationship between Emotional Stability and work performance ($\rho = .13$, $CVL = .01$, $CVU = .22$) across multiple worker types. Salgado (1998) found the relationship between Emotional Stability and performance was higher in a military, compared to a civilian, sample (military $\rho = .27$; civilian $\rho = .11$). As Salgado's sample was specifically European, we seek to test these results in an American context. We expect that Soldiers who tend to be anxious, moody, and negative will perform less effectively across a range of measures than those who tend to be calm, resilient, and optimistic.

Hypotheses

Based on this review, we proposed the following hypotheses:

- *Research Question: Do Big Five trait scores, as captured by TAPAS, predict Soldier outcomes?*
 - *Hypothesis 1: Conscientiousness will show a small relationship with Soldier performance.*
 - *Hypothesis 2: Emotional Stability will show a very small relationship with Soldier performance.*

Note that two constructs used to measure performance, Counterproductive Soldier Behavior and Motivation to Lead - Calculative are negatively keyed, and negative relationships are indicative of higher performance.

Method

Data for Analysis

This research effort used a subset of the Cycle 21 VAST database (see Chapter 2). We summarize below the various scales available on both the predictor and criterion side.

Tailored Adaptive Personality Assessment System (TAPAS)

The predictors for this research effort were the various personality facets that are part of TAPAS. Generally, these facets can be grouped according to the Big Five constructs of Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (Emotional Stability). Not all facets fit cleanly into this grouping. Some span between certain constructs, while others are of special import to the Army and were designed to predict outcomes that might not typically be captured by the Big Five dimensions. Given that most research on the relation between personality and performance is conducted with the Big Five, we constructed composites of TAPAS facets that align with the Big Five, as feasible, to allow for evaluation of relations that commonly appear in the literature. While we had hoped to examine the underlying factor structure of TAPAS, the pattern of missing data associated with the various versions of TAPAS on which data have been collected led us to use a content-based strategy to derive approximations of two Big Five factor scores.

Criterion Data

The VAST database has various criterion measures that assess different dimensions of job performance. These include job knowledge tests (JKTs), the Army Life Questionnaire (ALQ), and performance ratings (PRS). Although the JKTs index performance, we do not expect them to be related to the personality constructs measured by the TAPAS (see McCloy, Campbell, & Cudeck, 1994), so we did not include them in these analyses. While a brief overview of the ALQ and PRS measures is provided below, additional details on these measures can be found in Chapter 3 and in Knapp et al. (2020).

The ALQ contains a variety of self-report scales that are of interest to this investigation. Given the focus on the prediction of performance criteria, the contextual performance behaviors and attitudes were used as criteria. These included Counterproductive Soldier Behavior (CSB), Organizational Citizenship Behavior/Leadership (OCB/L), Resilience, and Motivation to Lead (MTL) with its three dimensions of Affective, Calculative, and Social-normative.

In addition to the self-report criteria captured by the ALQ, the use of performance ratings has also been a core focus of the VAST research effort. These PRS scales are desirable because they can comprehensively cover several dimensions of performance that are not limited to core performance behaviors such as technical proficiency. These scales were developed to capture motivational and non-cognitive performance behaviors and have the additional benefit of not being self-report measures.

As noted in Chapter 3 supervisor performance ratings (PRS-S) are collected from Soldiers in-unit, and peer performance ratings (PRS-P) are collected from peers in training and in units. Cadre ratings were previously collected from Soldiers in IMT, but were discontinued when the PRS-P was introduced in 2018 and therefore were not included in the validation analyses reported in Chapter 4. We did, however, include cadre ratings in the analyses reported in this chapter. For all the rating scales, we used an Overall Performance composite score that was calculated similarly across the four measures (IMT PRS-P, in-unit PRS-P, Cadre PRS, and PRS-S). The Cadre PRS Overall Performance composite score is based on ratings on five dimensions and overall performance made on a 5-point scale. More details on this measure are provided in Knapp et al. (2020).

Analytic Approach

The 71 MOS included in the VAST database cover a broad range of occupations critical to the Army's mission. Given the focus on non-cognitive performance, job-specific variance was expected to be small; however, we used multilevel modeling (MLM) to account for it where needed. Using MLM accounts for variance between MOS that would otherwise be lost if all MOS were treated as equivalent.

The characteristics of the data are well suited to this approach given that we have a sufficient sample size at both level 1 (Soldier) and level 2 (MOS). To ensure that sample sizes were sufficient, we followed best practices regarding estimating statistical power in MLM (Scherbaum & Ferreter, 2009). In instances where the level 1 and/or level 2 sample size was insufficient, we used ordinary least squares regression to model these criteria. MLM has not been used in other VAST analyses, but this investigation could inform the type of routine validation work presented in Chapter 4. Furthermore, given that the routine validation analyses focus on cognitive performance and attrition in addition to contextual performance, there may be more job-specific variance than that found here.

MLM accounts for criterion variance that occurs at level 2 (MOS). To start, we calculate indices to quantify the amount of variance that occurs at level 2. These indices are the intraclass correlation (ICC) and the design effect (Peugh, 2010). To illustrate, consider Conscientiousness to be the predictor and overall job performance to be the criterion for the following equations. To start, we estimate an unconditional means (i.e., random effect) model using the below equations:

$$\text{Level} - 1: Y_{ij} = \beta_{0j} + r_{ij} \quad (1)$$

$$\text{Level} - 2: \beta_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

In equation 1, the performance score of Soldier i in MOS j (Y_{ij}) is modeled as a function of the mean performance score for MOS j (β_{0j}) plus a residual term reflecting individual Soldier differences around the mean of MOS j (r_{ij}). In equation 2, the performance mean for MOS j (β_{0j}) is modeled as a function of a grand-mean performance score (γ_{00}) plus an MOS-specific deviation from the grand mean (u_{0j}). Substituting the right-hand side of equation 2 for β_{0j} in equation 1, yields the combined unconditional MLM equation:

$$\text{Combined: } Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}. \quad (3)$$

Fitting this model yields two estimates needed to calculate the ICC. The level 1 variance, σ^2 , describes the variation in performance across Soldiers within an MOS, and the level 2 variance, τ_{00} , describes the variation in performance across MOS. Using these estimates, the ICC is calculated as:

$$ICC = \tau_{00} / (\tau_{00} + \sigma^2). \quad (4)$$

Literature reports ICCs between .05 and .20 being common in cross-sectional MLM (Peugh, 2010), although this alone is not sufficient to warrant a MLM for nested data. Additionally, the design effect quantifies the effect of independence violations on standard error estimates and serves as a multiplier that is applied to standard errors to correct for negative bias resulting from nested data. The design effect is calculated as:

$$\text{Design Effect} = 1 + (n_c - 1)ICC \quad (5)$$

where n_c is the harmonic mean of the number of Soldiers per MOS and ICC is the value calculated in equation 4. Although there is no firm standard, researchers recommend that MLM be used when the design effect is greater than 2.0 (Muthen, 1994).

Forming Predictor Composites

Facets for use in the Big Five composites were informed by past research aligning the TAPAS facets with corresponding Big Five facets (Drasgow et al., 2012). We examined factor structures for the facets using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) to ensure alignment with past research. As described in further detail in the results, unit-weighted composites guided by literature-based groupings were used due to poor structure and fit indices for factor analytic models. Without item level data available, examination of why the factor analytic models are poor was limited and modifications were not possible due to the low number of facets available for each composite.

Results

The VAST Cycle 21 data file has notable sparseness with regard to TAPAS facets due to historical variation in facet inclusion across TAPAS versions (see Chapter 2). We identified six facets with sufficient overlap across TAPAS administrations and reasonably consistent sample sizes from which to compute higher-order composites: Achievement, Order, Non-Delinquency, Adjustment, Even Tempered, and Optimism. Given these six facet scores, trait scores could be computed for just two of the Big Five factors: Conscientiousness and Emotional Stability. The TAPAS facets for Conscientiousness were Achievement, Order, and Non-Delinquency, and the facets for Emotional Stability were Adjustment, Even Tempered, and Optimism. Note also that three of six facets typically associated with Conscientiousness (Responsibility, Self-Control, and Virtue) were not included in our Conscientiousness score due to limited inclusion in prior TAPAS versions. These three facets of Emotional Stability are unlikely to cover the trait's full construct space as well. Goldberg (1999) notes six facets of Emotional Stability of Anxiety, Anger, Depression, Self-Consciousness, Immoderation, and Vulnerability in the International Personality Item Pool Big Five.

The TAPAS trait analysis sample reported here is a subset of the IMT and in-unit samples described in Chapter 2. The trait sample comprised 10,129 cases, which are those Soldiers with valid data from the six target TAPAS facets and one or more non-administrative criterion measures.

Descriptive statistics for the six target TAPAS facets and two Big Five composites in this sample are presented in Table 5.1. Correlations between facets and personality scales are presented in Table 5.2. Correlation coefficients among Conscientiousness facets, ranging from .11 to .17 (all $p < .01$) were slightly lower than among Emotional Stability facets, ranging from .19 to .27 (all $p < .01$). Correlations between traits and their respective facets were all above .60 ($p < .01$), and below .19 for all cross-trait correlations, indicating clear convergent and discriminant validity. Although discriminant validity correlations are significant at $p < .05$, this is due to sample size and should not detract from the much smaller magnitude compared to the correlations between traits and their facets.

Table 5.1 TAPAS Facets and Trait Composites Descriptive Statistics

Scale	<i>M</i>	<i>SD</i>	Min	Max
<i>Facet</i>				
Achievement	0.03	0.98	-2.33	2.33
Order	-0.06	0.97	-2.33	2.33
Non-Delinquency	0.05	0.98	-2.33	2.33
Adjustment	0.06	0.98	-2.33	2.33
Even Tempered	0.04	0.97	-2.33	2.33
Optimism	0.03	0.97	-2.33	2.33
<i>Trait Composite</i>				
Conscientiousness	0.01	0.64	-2.24	2.24
Emotional Stability	0.04	0.68	-2.33	2.33

Note. $N = 10,129$. TAPAS scores are standardized. Conscientiousness facets are Achievement, Order, and Non-Delinquency; Emotional Stability facets are Adjustment, Even Tempered, and Optimism.

For Soldiers who had data for all six TAPAS facets, descriptive statistics for criteria are presented in Table 5.3. For IMT criteria, sample sizes ranged from 1,243 for PRS-P ratings, to 6,703 cases for Cadre PRS. In-unit criteria had somewhat fewer cases, ranging from 193 for PRS-P to 1,591 for PRS-S. Correlations among criterion scales are presented in Table 5.4.

Table 5.2. Correlations Among Predictor Measures

	1	2	3	4	5	6	7
1. Achievement	--						
2. Order	.15**	--					
3. Non-Delinquency	.17**	.11**	--				
4. Adjustment	.10**	-.07**	.01	--			
5. Even Tempered	.11**	.00	.19**	.22**	--		
6. Optimism	.17**	-.01	.10**	.27**	.19**	--	
7. Conscientiousness	.68**	.64**	.65**	.02*	.15**	.13**	--
8. Emotional Stability	.18**	-.04**	.15**	.72**	.68**	.70**	.15**

* $p < .05$. ** $p < .01$. Conscientiousness facets are Achievement, Order, and Non-Delinquency; Emotional Stability facets are Adjustment, Even Tempered, and Optimism.

Table 5.3. Criterion Scale Descriptive Statistics for Analysis Sample

Scale	<i>n</i>	<i>M</i>	<i>SD</i>	Min	Max
<i>IMT</i>					
ALQ Counterproductive Soldier Behavior	1,472	1.89	0.65	1	5
ALQ Motivation to Lead-Affective	1,471	3.50	0.84	1	5
ALQ Motivation to Lead-Noncalculative	1,472	2.29	0.86	1	5
ALQ Motivation to Lead-Social-normative	1,472	4.04	0.73	1	5
ALQ Organizational Citizenship Behavior/Leadership	1,472	3.56	0.69	1	5
Peer Performance Ratings Overall Performance Composite	1,243	3.51	0.62	1	5
Cadre Performance Ratings Overall Performance Composite	6,703	4.31	1.17	1	7
Resilience	1,472	4.07	0.61	1	5
<i>In-Unit</i>					
ALQ Counterproductive Soldier Behavior	479	1.92	0.57	1	4
ALQ Motivation to Lead-Affective	479	3.78	0.81	1	5
ALQ Motivation to Lead-Noncalculative	479	2.11	0.86	1	5
ALQ Motivation to Lead-Social-normative	479	4.10	0.70	1	5
ALQ Organizational Citizenship Behavior/Leadership	479	3.61	0.73	1	5
Peer Performance Ratings Overall Performance Composite	193	3.87	0.71	1	5
Supervisor Performance Ratings Overall Performance Composite	1,591	5.10	1.12	1	7
Resilience	479	3.96	0.61	1	5

Table 5.4. Criterion Score Intercorrelations

	CSB	MTL-A	MTL-N	MTL-S	OCB/L	PRS-P	PRS-S	Resilience
CSB		-.18*	.31*	-.27*	-.23*	-.15*	-.04	-.29*
MTL-A	-.13*		-.22*	.67*	.50*	.19*	.19	.41*
MTL-N	.39*	-.08*		-.34*	-.29*	-.13	-.08	-.17*
MTL-S	-.31*	.62*	-.26*		.59*	.22*	.12	.51*
OCB/L	-.16*	.42*	-.25*	.50*		.17*	.22*	.50*
PRS-P	-.04	.10*	-.05	.08*	.10*		.17	.21*
PRS-S	-	-	-	-	-	-		.10
Resilience	-.29*	.39*	-.15*	.54*	.10*	.10*	-	

Note. IMT results appear below diagonal; In-unit results appear above diagonal. Cadre ratings are not presented here as they were removed from administration at about the same time that these scales were put into administration. CSB = Counterproductive Soldier Behavior, MTL-A = Motivation to Lead-Affective, MTL-N = Motivation to Lead-Calculative, MTL-S = Motivation to Lead-Social-normative, OCB/L = Organizational Citizenship Behaviors/Leadership, PRS-P = Performance Rating Scale-Peer, PRS-S = Performance Rating-Supervisor. * = $p < .05$.

Predictor Composites

As noted above, the sparseness of available data limited the number of composites that we were able to create. As such, we focused on Conscientiousness and Emotional Stability and chose facets for these that were best represented in the analysis database. For Conscientiousness, these were Achievement, Order, and Non-Delinquency, and for Emotional Stability, these were Adjustment, Even Tempered, and Optimism. To start, we estimated a two-factor EFA with the above six facets. The EFA results showed that the facets did adhere to the Big Five mapping as laid out in Drasgow et al. (2012). There was a mean item complexity of 1.3. Next, we fit a two-factor CFA. The model fit was poor with $\chi^2(8) = 36,898.56$, $p < .05$, $CFI = .82$, $TLI = .67$, $RMSEA = .09$, and $SRMR = .05$. Factor loadings for the EFA and CFA are shown in Table 5.5. Given the model results, we opted not to create composites using factor loading weights, but rather created unit-weighted composites by taking the mean of the three facets for each composite.

Table 5.5 Factor Analysis Results

Scale	EFA Loadings		CFA Loadings	
	Emotional Stability	Conscientiousness	Emotional Stability	Conscientiousness
Adjustment	.66	-.05	.49	
Even Tempered	.32	.25	.45	
Optimism	.40	.16	.50	
Achievement	.08	.41		.55
Order	-.17	.34		.25
Non-Delinquency	-.01	.51		.39

Note. EFA = exploratory factor analysis and CFA = confirmatory factor analysis.

Model Fitting

While preparing the data for MLM, we determined that many MOS for various criteria did not have sufficient sample sizes to be included in an MLM. A minimum level 1 sample was set at 20 cases, and any MOS below this threshold was omitted from consideration. This resulted in from 16-20 MOS being used in the MLM analyses. After applying the level 1 sample size filter, some criteria had fewer than five MOS remaining which is too few to fit an MLM. For these criteria, we treated the MOS as equivalent and modeled them with ordinary least squares regression. MLMs were fitted for all IMT criteria and the in-unit PRS-S while the remaining in-unit criteria were fitted with OLS. Power analyses were conducted for all criteria and was deemed to be sufficient. Given that only main effects were estimated in the following models, sample size requirements were lower than if interaction effects had been estimated. Even with the cases lost in listwise deletion of the data the size of the model samples was still sufficient.

IMT Models

Level 2 variance results for criteria fitted with MLMs are shown in Table 5.6. With the exception of the IMT Cadre ratings, level 2 variance as indexed by the ICC and design effect was modest. While these indexes are small, using OLS still allows for the possibility of standard errors to be biased (Aguinis, Gottfredson, & Culpepper, 2013). In contrast to the others, the ICC and design effect for the IMT Cadre ratings was relatively large suggesting that these performance ratings contribute a significant amount of variability, and MLM is certainly necessary for this criterion. Lastly, we observed design effects greater than 2.0 for the IMT PRS-P and MTL Social-normative, suggesting that independence violations do exist in these criteria despite the low ICCs.

Table 5.6 Level 2 Variance Results for Multilevel Models

Criteria	n_2	ICC	Design Effect
<i>IMT</i>			
Cadre PRS	17	.239	15.90
PRS-P	16	.071	3.89
Motivation to Lead-Affective	17	.005	1.21
Motivation to Lead-Calculative	17	.019	1.85
Motivation to Lead-Social-normative	17	.029	2.27
Organizational Citizenship Behaviors/Leadership	17	.006	1.25
Counterproductive Soldier Behavior	17	.018	1.76
Resilience	17	.019	1.69
<i>In-Unit</i>			
PRS-S	21	.000	1.00

Note. n_2 = Level 2 sample size; ICC = intraclass correlation, PRS-P = Performance Rating Scale-Peer, PRS-S = Performance Rating-Supervisor

Parameter estimates for the IMT MLM models are shown in Table 5.7. The results support Hypotheses 1 and 2. Significant positive relationships for Conscientiousness were shown for the IMT scales of Motivation to Lead-Affective ($\gamma = .10, t = 2.98, p < .05$), Motivation to Lead-Social-normative ($\gamma = .07, t = 2.25, p < .05$), and Organizational Citizenship Behaviors/Leadership ($\gamma = .08, t = 2.69, p < .05$). A significant negative relationship was shown

for Conscientiousness and Counterproductive Soldier Behavior in IMT ($\gamma = -.09, t = -3.45, p < .05$). For Emotional Stability, significant positive relationships were shown for Organizational Citizenship Behavior/Leadership ($\gamma = .08, t = 2.91, p < .05$) and Resilience ($\gamma = .07, t = 3.12, p < .05$), and a negative relationship with Counterproductive Soldier Behavior ($\gamma = -.05, t = -1.96, p < .05$). As hypothesized, these effects were small and the effects for Emotional Stability were generally smaller than those of Conscientiousness, with the exception of the Resilience criterion.

Table 5.7 IMT Model Results

Criteria	df	Pseudo- R^2	Predictor	Estimate	SE	t
<i>IMT</i>						
Cadre	6684	.07	Conscientiousness	.03	.02	1.44
			Emotional Stability	.02	.02	1.16
PRS-P	1224	.07	Conscientiousness	.02	.03	0.81
			Emotional Stability	-.04	.03	-1.45
MTL-A	1446	.01	Conscientiousness	.10	.03	2.98*
			Emotional Stability	-.01	.03	-0.34
MTL-N	1453	.03	Conscientiousness	-.06	.04	-1.76
			Emotional Stability	-.06	.03	-1.81
MTL-S	1454	.03	Conscientiousness	.07	.03	2.25*
			Emotional Stability	.00	.04	0.13
OCB/L	1455	.02	Conscientiousness	.08	.03	2.69*
			Emotional Stability	.08	.03	2.91*
CSB	1453	.03	Conscientiousness	-.09	.03	-3.45*
			Emotional Stability	-.05	.03	-1.96*
Resilience	1456	.03	Conscientiousness	.04	.03	1.53
			Emotional Stability	.07	.02	3.12*

Note. PRS-S = Performance Rating Scale-Supervisor, PRS-P = Performance Rating Scale-Peer, MTL-A = Motivation to Lead-Affective, MTL-N = Motivation to Lead-Calculative, MTL-S = Motivation to Lead-Social-normative, OCB/L = Organizational Citizenship Behaviors/Leadership, CSB = Counterproductive Soldier Behavior, * = $p < .05$

In-Unit Models

Parameter estimates for the In-Unit PRS-P MLM and OLS models for other criteria are shown in Table 5.8. Hypothesis 1 was supported by these results while Hypothesis 2 was partially supported. Conscientiousness had significant positive relationships with In-Unit Motivation to Lead – Affective ($\beta = .11, t = 1.97, p < .05$), Organizational Citizenship Behavior/Leadership ($\beta = .11, t = 2.03, p < .05$), and Resilience ($\beta = .09, t = 2.16, p < .05$) while there was a significant negative relationship with In-Unit Counterproductive Soldier Behavior ($\beta = -.24, t = -6.13, p < .01$). Emotional Stability did not show any significant positive relationships; however, there was a significant negative relationship with Motivation to Lead – Calculative ($\beta = -.13, t = -2.18, p < .05$). As with the MLMs, the significant effects for these models were generally small.

Table 5.8 In-Unit Model Results

Criteria	<i>df</i>	<i>R</i> ²	Predictor	Estimate	<i>SE</i>	<i>t</i>
<i>MLM</i> ¹						
PRS-S	1588	.00	Conscientiousness	.10	.05	2.25*
			Emotional Stability	.03	.04	0.72
<i>OLS</i>						
PRS-P	190	.01	Conscientiousness	.07	.08	0.75
			Emotional Stability	-.09	.08	-1.18
MTL-A	476	.01	Conscientiousness	.11	.06	1.97*
			Emotional Stability	-.06	.06	-1.08
MTL-N	476	.02	Conscientiousness	-.11	.06	-1.74
			Emotional Stability	-.13	.06	-2.18*
MTL-S	476	.00	Conscientiousness	.05	.05	0.96
			Emotional Stability	-.01	.05	-0.29
OCB/L	476	.01	Conscientiousness	.11	.05	2.03*
			Emotional Stability	.00	.05	0.08
CSB	476	.07	Conscientiousness	-.24	.04	-6.13***
			Emotional Stability	.04	.04	0.95
Resilience	476	.01	Conscientiousness	.09	.04	2.16*
			Emotional Stability	.02	.04	0.35

Note. ¹*R*² for the MLM model is a Psuedo-*R*². PRS-P = Performance Rating Scale-Peer, MTL-A = Motivation to Lead-Affective, MTL-N = Motivation to Lead-Calculative, MTL-S = Motivation to Lead-Social-normative, OCB/L = Organizational Citizenship Behaviors/Leadership, CSB = Counterproductive Soldier Behavior, * = *p* < .05, *** = *p* < .001

Discussion

Relations among Personality Trait Scores and Contextual Performance

This investigation sought to test whether relations between broad personality trait scores and selected criterion measures are consistent with those found in the literature and whether these relationships would persist across occupations after correcting for MOS-specific clustering. In spite of the limitations of the dataset, our findings have several key implications. Conscientiousness predicted ALQ scales Motivation to Lead-Affective for MLM and OLS, Motivation to Lead-Social-normative for MLM, OCB/L for MLM and OLS, and Counterproductive Soldier Behavior for MLM and OLS. While these findings are significant, the magnitude of the relationships, as indexed by the *R*², were notably smaller than those found by Barrick et al. (2001), Van Aarde et al. (2017), and others. The reasons for this may lie in the measurement properties of the Conscientiousness and Emotional Stability composites. First, neither composite captured the full breadth of the Big Five traits, so there is likely meaningful construct space that is being missed in the composite scores. Second, referring to Table 5.5, the factor loadings of the TAPAS facets on the Big Five factor were quite low. This likely limits the reliability of the composites themselves which would in turn attenuate relationships with criteria. Other factors differentiating the present investigation with prior meta-analyses is the uniformly high-stakes setting in which personality was assessed, as well as the TAPAS measurement model.

While the investigation here was limited by the lack of completeness in TAPAS facet scores, this will be somewhat ameliorated as more data are collected. The newest TAPAS versions currently in use have a common 13 facets assessed for all Soldiers while experimental facets are administered in a separate section. With continued data collection, more complete data will be obtained to estimate Big Five composites. While attempting to construct all five Big Five composites out of 13 facets would be tenuous, constructing more than two composites may be possible. As development of the TAPAS continues, experimental facets will be administered to more Soldiers allowing for a broader examination of personality and non-cognitive performance.

Application of Multilevel Modeling (MLM)

As noted previously, it is worth exploring the application of MLM to the TAPAS composite validation analyses described in Chapter 4. Sample sizes will be larger given more complete data which ideally would allow for examination of more criteria than was possible here. Despite the mostly low ICC values reported here, relatively small ICC estimates (e.g., .05, Peugh, 2010) can bias standard errors from OLS models. Furthermore, even in the presence of low level 2 variance estimates, an MLM model is not incorrect when the data is nested. While an OLS model may be more parsimonious, MLM models should be estimated to assess the heterogeneity in level 2 units. If the heterogeneity is small or nonexistent then OLS models may be used, but without estimating the MLM models, heterogeneity that may be present would remain undetected leading to biased standard errors for OLS.

Factoring in Occupational Characteristics

Previous research on person-job fit (e.g., Judge & Zapata, 2015) suggests that personality-performance relationships may depend on the characteristics of a given job. We believe it would be valuable for future research to explore differences in personality-performance relationships across occupations. Information on MOS demand characteristics could enhance the predictive value of personality assessments for Army jobs. Some progress has been made in measuring job demands. For example, the Department of Labor's *My Next Move for Veterans* has extrapolated civilian occupation characteristics to military occupations (McCloy, 1994). However, this system omits some of the most common Army MOS, such as Infantry, for which there is less basis for comparison to civilian occupations. Several new Army programs may soon provide this information. The Officer Assignment Interactive Module Version 2 (AIM 2.0) and the Integrated Personnel and Pay System – Army (IPPS-A) introduce standardized job requirement ratings. Once these systems are developed, interactive effects between MOS characteristics and Soldier personality may be more readily examined.

Summary

In this chapter, we explored the feasibility of computing traditional five-factor model personality scales from TAPAS facets. We re-analyzed TAPAS data to construct scale scores from TAPAS facets using theory-driven, unit-weighted composites since factor analysis of facet scores yielded poor model fit.

Multilevel modeling provides a means to identify relationships when data show clustering which can distort results of OLS regression. We applied MLM to identify meaningful clustering in criterion measures within MOS, which we statistically controlled for in our models.

We regressed multiple performance criteria on Conscientiousness and Emotional Stability scale scores. We found that Conscientiousness and Emotional Stability composites, computed from TAPAS facets, predicted performance-related variables in a manner generally consistent with prior research in that most relations were statistically significant. However, the magnitude of the relations were lower than expected. Several possible reasons for these findings were described.

CHAPTER 6: REFLECTIONS

Deirdre J. Knapp (HumRRO) and Cristina D. Kirkendall (ARI)

Criterion-Related Validity Evidence

Thus far, the VAST biannual analyses and the TOPS IOT&E work that preceded VAST has focused on the incremental validity of TAPAS composite scores over AFQT scores to predict important outcomes. Recognizing that there are multiple training and in-unit performance outcomes of interest to policymakers, this research incorporates criterion measures covering job knowledge, perceptions of fit and other self-report indicators, performance ratings, and continuance behavior. Biannual reports routinely present results related to attrition, and reenlistment behavior has been examined as well (Hughes, O'Brien, Reeder, & Purl, 2020; Rogers, Hughes, Marshall, Moore, & Kirkendall, 2021). Particularly as the analysis sample sizes have increased over the years, the basic results – indicating that TAPAS Will-Do and Adaptation scores improve prediction of outcomes that AFQT was not designed to predict – have been quite stable from one six-month cycle to the next. The degree of improvement varies by outcome, but even small increments can be valuable in a selection system that processes tens of thousands of applicants annually.

Despite the large sample sizes, the TAPAS data are rather sparse, especially when combined with non-administrative criterion measures. As illustrated in Chapter 5, only six of the 26 facets that TAPAS has been used to measure have been common across all TAPAS versions. Moreover, the latest formulations of the three selection composites of interest to the Army cannot be applied to most of the currently available TAPAS cases and a fourth composite can only be calculated for those who have taken the TAPAS versions introduced at the end of 2019. For the time being, therefore, we need to combine the old and new formulations of the three TAPAS composites and hold off criterion-related validation of the fourth until we have criterion data on a sufficient number of cases. That will take longer than would ordinarily be expected due to the COVID-19 worldwide pandemic which has prevented collection of new non-administrative criterion data for over a year.

Evidence of incremental validity over AFQT is of primary interest because it reflects the operational context in which TAPAS is expected to be used for Army enlisted personnel selection decisions. We examined criterion-related validity of the TAPAS alone in the course of using meta-analysis to determine whether different TAPAS versions yield varying results (Hughes, 2020), and those results suggested that the different versions yield comparable results. In the present report, we have also looked at criterion-related validity associated with different configurations of facet scores intended to replicate two of the Big Five personality factors. In that effort, we also factored in the possibility of MOS differences systematically affecting our validity estimates through the use of multilevel modeling. Our future work will continue to explore ways to improve the accuracy of both the absolute and incremental validity estimates computed using the VAST database. Moreover, as the Army considers specific policies for operational decision-making, we can inform and evaluate those policies by simulating them using the VAST database.

Other TAPAS Evaluation Efforts

Whereas the TOPS IOT&E and VAST research programs have focused primarily on evidence of criterion-related validity, there have been and continue to be examinations of the TAPAS that use a broader lens for evaluating its suitability for military selection and classification decision-making. These include considerations related TAPAS' design, development, derivation of composite and cut scores, administration, retest policies, and so forth.

As the TOPS IOT&E was getting underway, ARI convened a panel of independent experts to critique the TAPAS as well as the IOT&E research plan. The review was positive about the promise of TAPAS and soundness of the research plan. Suggestions included the need for more documentation of how TAPAS works, its underlying factor structure, and other psychometric details which would allow a more complete evaluation. Recommendations related to the IOT&E method suggested minor changes. In light of the evolution of the research program in the past 10 years, we plan to take a fresh look at those recommendations to see if they prompt ideas for adjustments to our work moving forward.

Promising criterion-related validation results reported by the Army generated joint-service interest in the TAPAS. The Defense Testing and Assessment Center (DTAC) sponsored another expert panel review of TAPAS (Roberts et al., 2019). This latest panel focused entirely on the question of TAPAS' readiness for operational use. Its critique was guided by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The panel offered detailed recommendations in several areas, including those that had also been offered by the earlier expert panel review (e.g., need for more information about TAPAS' design, development, and psychometric properties).

Partly in response to recommendations made by Roberts et al. (2019) DTAC also sponsored development of a *validity argument* to support the use of TAPAS for military selection and classification decision-making which has recently been completed (Ford, Knapp, Russell, & Thacker, 2022). This approach to thinking about validity originated in the educational testing arena (e.g., Bennett, 2010), though it has recently been applied to the AFQT (Thacker et al., 2020) and is currently being applied to ASVAB's use in classification decisions. This method involves specifying a *theory of action* which is a succinct statement of why an assessment is expected to be useful for its intended purpose. That becomes the basis for an *interpretive argument* which specifies claims and associated assumptions about the assessment and how it is used. Compilation (and/or generation) of evidence related to those assumptions then forms the basis of the validity argument. This work is expected to inform the research and development needed to create new versions of TAPAS geared to selection and classification needs across all services.

What's Next for TAPAS?

ARI is researching the use of TAPAS in a number of ways other than as an across-the-board qualification requirement for selection into the Army. As called for in the Army People Strategy implementation plan, there is currently a pilot program to count up to 5% AFQT Category IIIB applicants as Category IIIA for accessioning purposes, provided those applicants score sufficiently high on TAPAS. ARI is also looking into TAPAS as a potential component (alone or in combination with ASVAB, the Adaptive Vocational Interest Diagnostic [AVID], and/or other scores) for determining MOS fit to support classification decisions (e.g., Bynum, Trippe, Diaz, & Ingerick, 2013; Nye et al., 2012, 2020) and has initiated research to develop occupation-specific TAPAS composites that could be used to qualify applicants for select MOS (e.g., Knapp, Wind, Kiger, Posnock, & Taylor, 2021).

With regard to the instrument itself, ARI is also conducting research to examine additional facets that could be included on TAPAS and to improve mechanisms for detecting attempted response distortion on TAPAS. Efforts are also underway to determine what facets would be most suitable for a joint-service version of TAPAS.

It bears mentioning that ARI is also using in-service variations of TAPAS to make post-enlistment selection decisions. In particular, the NCO Special Assignment Battery (NSAB) has been investigated for use in selection of Special Operations Forces positions, recruiters, drill sergeants, instructors, and senior NCO leadership positions (e.g., Horgen et al., 2013; Nye et al., 2014).

In part as a result of feedback from the TAPAS Evaluation Panel (Roberts et al., 2019), both ARI and the U.S. Air Force have initiated efforts to more fully document details related to TAPAS' design and psychometric characteristics. Even as there is additional work to be done for TAPAS and its associated documentation to more fully meet professional standards for high stakes assessments, there continues to be considerable promise for its use in selection of enlisted personnel and other applications. Based on the criterion-related validity evidence presented here, there is clearly merit to continued exploration of TAPAS as a tool for adding personality characteristics to a system that currently relies primarily on cognitive ability and physical fitness for selection and classification decision-making.

REFERENCES

- Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management* 39(6), 1490-1528.
- Allen, M. T., Knapp, D. J., & Owens, K. S. (2016). *Validating future force measures (Army Class): Concluding analyses* (Technical Report 1355). U.S. Army Research Institute for the Behavioral and Social Sciences.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1), 1-26.
- Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and assessment*, 9(1-2), 9-30.
- Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, 8(2-3), 70-91.
- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, 71(3), 460-502.
- Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, 18(1), 36-52.
- Bynum, B. H., & Mullins, H. M. (Eds.) (2017). *Tier One Performance Screen initial operational test and evaluation: 2013 annual report* (Technical Report 1359). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Bynum, B. H., Trippe, M., Diaz, T., & Ingerick, M. (2013). Evaluation of TAPAS potential for classification. In D.J. Knapp & C.D. Kirkendall (Eds.) *Tier One Performance Screen initial operational test and evaluation: 2012 interim report* (Technical Report 1332). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Campbell, J. P., & Knapp, D. J. (Eds.) (2001). *Exploring the limits in personnel selection and classification*. Lawrence Erlbaum Associates, Inc.
- Campbell, J. P., Hanson, M. A., & Oppler S. H. (2001). Modeling performance in a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 307-334). Lawrence Erlbaum Associates, Inc.

- Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations* (pp. 35-70). Jossey-Bass.
- Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology*, 43, 313-333.
- Carretta, T. R., Ree, M. J., & Teachout, M. S. (2016). Training affects variability in training performance both within and across jobs. *International Journal of Selection and Assessment*, 24, 72-76.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd edition)*. Lawrence Erlbaum Associates, Inc.
- Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115, 928-935.
- DMDC (2008). *ASVAB Technical Bulletin No. 3 CAT-ASVAB Forms 5-9*. Defense Manpower Data Center.
- Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army selection and classification decisions* (Technical Report 1311). U.S. Army Research Institute for the Behavioral and Social Sciences. <https://discover.dtic.mil/ADA564422>
- Farmer, W. L., Watson, S. E., Alderton, D. L., Michael, P. G., & Hindelang, R. (2006). *Improving person-job congruence during the classification process: Item development and initial testing of a pictorial interest instrument* (NPRST-TN-06-8). Navy Personnel Research, Studies, and Technology Division, Bureau of Naval Personnel.
- Ford, L. A., Knapp, D. J., Russell, T. R., & Thacker, A. A. (2022). *Validity argument 1.0 for the Tailored Adaptive Personality Assessment System (TAPAS)* (2021 No. 095). Human Resources Research Organization.
- Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several Five-Factor models. *Personality Psychology in Europe*, 7, 7-28.

- Graves, C., & Taylor, W. (2020). Description and psychometric properties of objective criterion measures. In D.J. Knapp & C.D. Kirkendall (Eds.) *Tier One Performance Screen initial operational test and evaluation: Capstone report* (Technical Report 1380). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Horgen, K. E., Nye, C. D., White, L. A., LaPort, K. A., Hoffman III, R. R., Drasgow, F., Chernyshenko, O. S., Stark, S., & Conway, J. S. (2013). *Validation of the Noncommissioned Officer Special Assignment Battery* (Technical Report 1328). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Houston, J. S., Borman, W. C., Farmer, W. F., & Bearden, R. M. (2006). *Development of the Navy Computer Adaptive Personality Scales (NCAPS)* (NPRST-TR-06-2). Navy Personnel Research, Studies, and Technology Division Bureau of Naval Personnel.
- Hughes, M. G. (2018). Evidence for the predictive validity of the TAPAS. In D. J. Knapp & C. D. Kirkendall (Eds.) *Tier One Performance Screen initial operational test and evaluation: 2015-2016 biennial report* (Technical Report 1367). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Hughes, M. G. (2020). Evidence for the predictive validity of the TAPAS. In D.J. Knapp & C.D. Kirkendall (Eds.) *Tier One Performance Screen initial operational test and evaluation: Capstone report* (Technical Report 1380). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Hughes, M. G., O'Brien, E. L., Reeder, M. C., & Purl, J. (2020) Attrition and reenlistment in the Army: Using the Tailored Adaptive Personality Assessment System (TAPAS) to improve retention. *Military Psychology*, 32, 36-50.
- Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Intelligence*, 29, 340-362.
- Ingerick, M., Diaz, T., & Putka, D. (2009). *Investigations into Army enlisted classification systems: Concurrent validation report* (Technical Report 1244). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology*, 98(6), 875-925.
- Judge, T. A., & Zapata, C. P. (2015). The person–situation debate revisited: Effect of situation strength and trait activation on the validity of the Big Five personality traits in predicting job performance. *Academy of Management Journal*, 58(4), 1149-1179.
- Kirkendall, C. D., Nye, C. D., Rounds, J., Drasgow, F., Chernyshenko, O. S., & Stark, S. (2020). Adaptive Vocational Interest Diagnostic: Informing and improving the job assignment process. *Military Psychology*, 32(1), 91-100.

- Knapp, D. J., & Heffner, T. S. (Eds.) (2009). *Predicting Future Force Performance (Army Class): End of Training Longitudinal Validation* (Technical Report 1257). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Heffner, T. S. (Eds.) (2010). *Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Heffner, T. S. (Eds.) (2011). *Tier One Performance Screen initial operational test and evaluation: 2010 annual report* (Technical Report 1296). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Heffner, T. S. (Eds.) (2012). *Tier One Performance Screen initial operational test and evaluation: 2011 interim report* (Technical Report 1306). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., Heffner, T. S., & White, L. (Eds.) (2011). *Tier One Performance Screen initial operational test and evaluation: Early results* (Technical Report 1283). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Kirkendall, C. D. (Eds.) (2018). *Tier One Performance Screen initial operational test and evaluation: 2015-2016 biennial report* (Technical Report 1367). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Kirkendall, C. D. (Eds.) (2020). *Tier One Performance Screen initial operational test and evaluation: Capstone report* (Technical Report 1380). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & LaPort, K. (Eds.) (2013a). *Tier One Performance Screen initial operational test and evaluation: 2011 annual report* (Technical Report 1325). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & LaPort, K. (Eds.) (2013b). *Tier One Performance Screen initial operational test and evaluation: 2012 interim report* (Technical Report 1332). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & LaPort, K. (Eds.) (2014). *Tier One Performance Screen initial operational test and evaluation: 2012 annual report* (Technical Report 1342). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., Sager, C. E., & Tremble, T. R. (Eds.). (2005) *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (Technical Report 1168). U.S. Army Research Institute for the Behavioral and Social Sciences.

- Knapp, D. J., Salmon, E., Kirkendall, C. D., & Burgoyne, T. (2020). Description and psychometric properties of performance rating criterion measures. In D.J. Knapp & C.D. Kirkendall (Eds.) *Tier One Performance Screen initial operational test and evaluation: Capstone report* (Technical Report 1380). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Tremble, T. R. (Eds.) (2007). *Concurrent validation of experimental Army enlisted personnel selection and classification measures* (Technical Report 1205). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D.J., Wind, A., Kiger, T., Posnock, S.J., & Taylor, W. (2021). *Improving the prediction of performance in cyber training* (Research Note 2021-05). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Wolters, H. M. K. (Eds.) (2017). *Tier One Performance Screen initial operational test and evaluation: 2014 annual report* (Technical Report 1358). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Lee, P., Joo, S. H., & Lee, S. (2019). Examining stability of personality profile solutions between Likert-type and multidimensional forced choice measure. *Personality and Individual Differences*, 142, 13-20.
- Maier, M. H. (1993). *Military aptitude testing: The past fifty years* (DMDC Technical Report 93-007). Defense Manpower Data Center.
- McCloy, R. A. (1994). Predicting job performance scores for jobs without performance data. In B. F. Green & A. S. Mavor (Eds.), *Modeling cost and performance for military enlistment: Report of a workshop*.
- McCloy, R.A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology*, 79(4), 493-505.
- McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1), 81-90.
- Moriarty, K. O., Campbell, R. C., Heffner, T. S., & Knapp, D. J. (2009). *Validating future force performance measures (Army Class): Reclassification test and criterion development* (Research Product 2009-11). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Muthen, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, 22, 376-398.
- Nesbitt, C. K., Salmon, E., & Kirkendall, C. D. (2020). Army Life Questionnaire. In D. J. Knapp & C. D. Kirkendall (Eds.) *Tier One Performance Screen initial operational test and evaluation: Capstone report* (Technical Report 1380). U.S. Army Research Institute for the Behavioral and Social Sciences.

- Nye, C. D., Beal, S. A., Drasgow, F., Dressel, J. D., White, L. A., & Stark, S. (2014). *Assessing the Tailored Adaptive Personality Assessment System for Army Special Operations Forces personnel*. (Technical Report 1971). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., Drasgow, F., Chernyshenko, O. S., Stark, S., Kubisiak, U. C., White, L. A., & Jose, I. (2012). *Assessing the Tailored Adaptive Personality Assessment System (TAPAS) as an MOS qualification instrument* (Technical Report 1312). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nye, C. D., White, L. A., Drasgow, F., Prasad, J., Chernyshenko, O. S., Stark, S., & Kubisiak, U. C. (2020). *Non-cognitive tools for military occupational specialties qualification* (Technical Report 1387). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48*(1), 85-112.
- Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93*, 959-981. <https://doi.org/10.1037/0021-9010.93.5.959>
- Roberts, J. S., Arthur, W., Reckase, M., Sackett, P., & Zenisky, A. (2019). *TAPAS evaluation project report on the readiness of TAPAS for use in selection and classification of military applicants* (2019 No. 076). Human Resources Research Organization.
- Rogers, A. P., Hughes, M. G., Marshall, A., Moore, C., & Kirkendall, C. D. (in preparation). *Predicting reenlistment behavior in the regular Army*. U.S. Army Research Institute for the Behavioral and Social Sciences.
- Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human resource management review, 16*(2), 155-180.
- Rose, M. R., Manley, G. G., & Weissmuller, J. J. (2013). *Development of two- and three-factor classification models for Air Force Battlefield Airmen (BA) and related AFSs*. Air Force Personnel Center Strategic Research and Assessment Branch.
- Roulin, N., Krings, F., & Binggeli, S. (2016). A dynamic model of applicant faking. *Organizational Psychology Review, 6*(2), 145-170.
- Salgado, J. F., & Tauriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3-30.
- Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives. *The Five-Factor Model of Personality: Theoretical Perspectives*, 21-50.

- Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods* 12(2), 347-367.
- Sharp, M. A., Foulis, S. A., Redmond, J. E., Canino, M.C., Cohen, B. S., Hauret, K., Frykman, P. N., Pierce, J. R., Westrick, R. B., Pacheco, B. M., Gebhardt, D. L., Zambraski, E. J. (2018). *Longitudinal validation of the Occupational Physical Assessment Test (OPAT)*. (Report #18-05). U.S. Army Research Institute of Environmental Medicine.
- Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The Life Outcomes of Personality Replication Project. *Psychological Science*, 30(5), 711-727.
- Strickland, W. J. (Ed.) (2005). *A longitudinal examination of first term attrition and reenlistment among FY1999 enlisted accessions* (Technical Report 1172). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.
- Thacker, A. A., Russell, T. L., Knapp, D. J., & Gribben, M. A. (2020). *Validity argument 1.0 for the Armed Forces Qualification Test (AFQT)* (2020 No. 07). Human Resources Research Organization.
- Trent, J. D., Barron, L. G., Rose, M. R., & Carretta, T. R. (2020). Tailored Adaptive Personality Assessment Systems (TAPAS) as an indicator for counterproductive work behavior. Comparing validity in applicant, honest, and directed faking conditions. *Military Psychology*, 32(1), 51-59. <https://doi.org/10.1080/08995605.2019.1652481>
- Trippe, D. M., Caramagno, J. P., Allen, M. T., & Ingerick, M. J. (2011). Initial evidence for the predictive validity and classification potential of the TAPAS. In D. J. Knapp, T. S. Heffner, & L. White (Eds.) *Tier One Performance Screen initial operational test and evaluation: Early results* (Technical Report 1283). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Trippe, D. M., Moriarty, K. O., Russell, T. L., Carretta, T. R., & Beatty, A. S. (2014). Development of a cyber/information technology knowledge test for military enlisted technical training qualification. *Military Psychology*, 26 (3), 182-198.
- U.S. Army (2019, October). Army People Strategy. <https://people.army.mil/wordpress/wp-content/uploads/2019/10/The-2020-Army-People-Strategy-Final.pdf>
- U.S. Army (2020, October). *2020 Army People Strategy: Military implementation plan*. <https://people.army.mil/overview-2/implementation/>
- Van Aarde, N., Meiring, D., & Wiernik, B. M. (2017). The validity of the Big Five personality traits for job performance: Meta-analyses of South African studies. *International Journal of Selection and Assessment*, 25(3), 223-239.

- Van Iddekinge, C. H., Putka, D. J., & Sager, C. E. (2005). Attitudinal criteria. In D. J. Knapp & T. R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (Technical Report 1168) (pp. 89-104). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Wolters, H., White, L., Drasgow, F., Stark, S., Nye, C., & Chernyshenko, O.S. (2017, July). *TAPAS research update: Ongoing and future research*. Presentation given to the Defense Advisory Committee on Military Personnel Testing.
- Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single statement measures. *Organizational Research Methods*, 23(3), 569-590.

APPENDIX A

LIST OF FREQUENTLY USED ACRONYMS

AFQT	Armed Forces Qualification Test
AHRC	Army Human Resources Command
AIT	Advanced Individual Training
ALQ	Army Life Questionnaire
APFT	Army Physical Fitness Test Score
ASVAB	Armed Services Vocational Aptitude Battery
ATRRS	Army Training Requirements and Resources System
ATSC	Army Training Support Center
CSB	Counterproductive Soldier Behavior
DMDC	Defense Manpower Data Center
EEEM	Expanded Enlistment Eligibility Metrics
ICTL	Information, Communication, and Technical Literacy
IMT	Initial Military Training
IRR	Interrater reliability
JKT	Job knowledge test
MEPCOM	Military Entrance Processing Command
MOS	Military Occupational Specialty
MTL	Motivation to Lead
NCO	Noncommissioned officer
NPS	Non-prior service
OCB	Organizational citizenship behavior
PRS-P	Performance rating scales completed by peers
PRS-S	Performance rating scales completed by supervisors or cadre
RITMS	Resident Individual Training Management System
SME	Subject matter expert
TAPAS	Tailored Adaptive Personality Assessment System
TAPDB	Total Army Personnel Database
TOPS IOT&E	Tier One Performance Screen initial operational test and evaluation
WTBD JKT	Warrior tasks and battle drills job knowledge test

APPENDIX B

PREDICTOR MEASURE PSYCHOMETRIC PROPERTIES IN THE APPLICANT SAMPLE

Table B.1. Means and Standard Deviations for Selected TAPAS Facet and Composite Scores in the Applicant Sample

TAPAS Composite/Facet	15D-Static/v4/5/7/8 (n = 101,023-421,846) ^a		v9 (n = 267,855)		v10 (n = 267,515)		v11 (n = 269,890)		v12-14 (n = 34,184)	
	M	SD	M	SD	SD	SD	M	SD	M	SD
<i>Individual TAPAS Facets</i>										
Achievement	-0.01	0.98	-0.04	0.98	-0.06	0.98	-0.03	0.99	-0.06	0.96
Adjustment	-0.02	0.98	—	—	—	—	—	—	-0.02	0.95
Attention Seeking	-0.04	0.98	—	—	—	—	-0.03	0.96	0.02	0.96
Dominance	-0.02	0.98	-0.04	0.97	-0.08	0.97	-0.06	0.99	-0.04	0.95
Even Tempered	0.01	0.98	0.00	1.00	-0.03	1.00	0.02	1.00	-0.07	0.96
Intellectual Efficiency	-0.02	0.97	0.02	0.97	-0.01	0.95	0.03	0.96	-0.04	0.95
Non-Delinquency	0.03	0.98	—	—	-0.02	0.97	—	—	-0.06	0.96
Optimism	0.00	0.98	-0.03	0.99	-0.01	0.98	-0.03	1.00	-0.06	0.96
Order	0.01	0.98	-0.01	0.96	-0.02	0.97	0.01	0.97	-0.02	0.95
Physical Conditioning	0.01	0.98	-0.04	1.01	-0.04	0.98	-0.01	1.00	0.01	0.94
Sociability	-0.01	0.98	-0.01	1.00	-0.01	1.01	-0.02	0.99	0.01	0.95
Team Orientation	-0.01	0.99	—	—	—	—	0.01	0.99	-0.02	0.95
Tolerance	0.00	0.98	0.10	0.98	0.10	0.98	0.11	0.97	-0.05	0.95
<i>TAPAS Composites</i>										
Can-Do	99.63	21.05	99.88	19.63	99.25	19.63	100.01	19.58	99.41	19.75
Will-Do	99.81	19.97	98.84	20.18	98.56	19.57	99.14	20.79	99.53	19.61
Adaptation	100.22	19.84	99.56	19.89	99.47	19.79	100.02	21.07	99.54	19.64
Conduct	100.77	20.96	—	—	—	—	—	—	98.37	20.22

^a Sample sizes vary due to TAPAS facets not being administered in every version.

Table B.2. Correlations among the AFQT and TAPAS Facets in the Applicant Sample

TAPAS Scale	1	2	3	4	5	6	7	8	9	10	11	12	13
1. AFQT													
<i>Individual TAPAS Facets</i>													
2. Achievement	0.06												
3. Adjustment	0.12	0.11											
4. Attention Seeking	0.05	0.05	0.11										
5. Dominance	0.10	0.33	0.10	0.24									
6. Even Tempered	0.09	0.16	0.23	-0.04	0.01								
7. Intellectual Efficiency	0.35	0.31	0.19	0.09	0.30	0.14							
8. Non-Delinquency	-0.03	0.22	0.02	-0.14	-0.01	0.26	0.05						
9. Optimism	0.06	0.19	0.26	0.08	0.14	0.23	0.14	0.14					
10. Order	-0.17	0.22	-0.06	-0.04	0.09	0.05	0.10	0.15	0.04				
11. Physical Conditioning	0.05	0.22	0.06	0.09	0.19	-0.04	0.08	-0.04	0.06	0.07			
12. Sociability	-0.10	0.14	0.10	0.35	0.26	0.08	0.10	0.02	0.15	0.03	0.06		
13. Team Orientation	-0.10	0.11	0.05	0.14	0.13	0.12	0.00	0.06	0.10	0.04	0.07	0.25	
14. Tolerance	0.03	0.12	0.03	0.05	0.06	0.17	0.14	0.07	0.10	0.04	-0.04	0.19	0.12

Note. $n = 101,023 - 1,261,290$. Not all TAPAS facets were administered in every version; missing correlations indicate that the facets were not administered on the same version. All correlations greater than or less than .00 are statistically significant at $p < .01$ (two-tailed).

Table B.3. Group Differences on the AFQT and TAPAS by Gender in the Applicant Sample

Measure	Male		Female		<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
<i>AFQT</i>	56.14	22.56	49.75	21.44	-0.29
<i>Individual TAPAS Facets</i>					
Achievement	-0.05	0.99	0.05	0.96	0.11
Adjustment	0.04	0.97	-0.26	0.98	-0.31
Attention Seeking	-0.01	0.98	-0.28	0.96	-0.10
Dominance	-0.05	0.98	-0.11	0.97	0.01
Even Tempered	0.00	0.99	-0.06	0.96	-0.01
Intellectual Efficiency	0.01	0.97	0.15	0.98	-0.07
Non-Delinquency	-0.03	0.98	-0.19	1.00	0.19
Optimism	-0.03	0.98	-0.04	0.95	0.04
Order	-0.07	0.95	-0.01	1.00	0.29
Physical Conditioning	0.06	0.99	-0.05	0.94	-0.34
Sociability	-0.03	0.99	0.15	0.95	0.08
Team Orientation	0.03	0.98	0.01	0.99	-0.14
Tolerance	-0.01	0.97	0.21	0.99	0.34
<i>TAPAS Composites</i>					
Can-Do	100.91	19.97	95.66	19.25	-0.27
Will-Do	100.10	20.26	96.18	19.27	-0.20
Adaptation	101.67	19.86	93.74	19.83	-0.40
Conduct	99.22	20.82	103.29	20.45	0.20

Note. Females are the referent group. Sample size for Females range from 22,385 to 292,522; for Males 78,656 to 968,768. Significant Cohen's *d* values, based on an independent sample *t*-test between the group means are in bold ($p < .05$, two-tailed).

Table B.4. Group Differences on the AFQT and TAPAS by Race and Ethnicity in the Applicant Sample

Measure	White, Non-Hispanic		Black		<i>d</i>	Hispanic		<i>d</i>	Asian		<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		<i>M</i>	<i>SD</i>		<i>M</i>	<i>SD</i>	
<i>AFQT</i>	60.76	21.83	44.39	19.47	-.77	48.27	20.87	-0.58	59.65	23.89	-0.05
<i>Individual TAPAS Facets</i>											
Achievement	0.03	0.99	-0.06	0.95	-.09	-0.10	0.97	-0.13	-0.29	0.97	-0.32
Adjustment	0.06	0.99	-0.12	0.95	-.17	-0.12	0.96	-0.18	-0.24	0.95	-0.30
Attention Seeking	-0.07	0.99	0.07	0.95	.14	-0.03	0.94	0.04	-0.04	0.95	0.03
Dominance	-0.02	1.02	-0.01	0.90	.01	-0.11	0.95	-0.09	-0.32	0.94	-0.29
Even Tempered	0.02	1.01	0.00	0.98	-.03	-0.03	0.97	-0.06	-0.15	0.95	-0.17
Intellectual Efficiency	0.03	0.99	0.02	0.91	-.01	-0.09	0.94	-0.12	-0.18	0.96	-0.22
Non-Delinquency	0.01	0.99	0.06	0.96	.05	-0.02	0.95	-0.03	-0.12	0.93	-0.13
Optimism	0.01	0.99	-0.03	0.98	-.04	-0.04	0.97	-0.05	-0.19	0.97	-0.20
Order	-0.13	0.97	0.19	0.96	.33	0.08	0.93	0.22	0.13	0.92	0.27
Physical Conditioning	0.05	1.01	-0.13	0.95	-.18	-0.03	0.97	-0.08	-0.14	0.97	-0.18
Sociability	-0.03	1.02	0.06	0.96	.09	-0.02	0.97	0.01	-0.11	0.97	-0.08
Team Orientation	-0.04	0.99	0.02	0.98	.05	0.05	0.98	0.08	0.16	0.99	0.20
Tolerance	-0.04	1.01	0.17	0.93	.21	0.21	0.93	0.25	0.28	0.91	0.32
<i>TAPAS Composites</i>											
Can-Do	102.18	20.44	97.35	18.52	-.24	96.77	19.14	-0.27	94.30	19.35	-0.39
Will-Do	100.63	20.80	97.85	18.70	-.14	98.06	19.48	-0.13	93.78	19.63	-0.33
Adaptation	101.84	20.39	96.38	19.47	-.27	98.98	19.66	-0.14	97.28	19.57	-0.22
Conduct	99.95	21.29	99.74	20.39	-.01	100.88	19.87	0.04	102.36	20.22	0.11

Note. The Minority group is the referent group. Sample size for White, Non-Hispanic ranges from 54,643 to 667,864; Black from 25,251 to 298,166; Hispanic from 15,636 to 213,251, Asian from 3,986 to 59,633. Significant Cohen's *d* values, based on an independent sample *t*-test between the group means are in bold ($p < .05$, two-tailed).

APPENDIX C

CRITERION MEASURE PSYCHOMETRIC PROPERTIES

Table C.1. Correlations among the ALQ Scores in the Full IMT and In-Unit Criterion Samples

Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Affective Commitment		0.67	0.40	0.48	0.57	0.55	0.30	-0.08	0.43	0.36	-0.22	0.40	0.02	0.06
2. Army Fit	0.71		0.44	0.50	0.60	0.59	0.27	-0.23	0.43	0.36	-0.35	0.46	-0.03	0.09
3. MOS Fit	0.45	0.49		0.64	0.28	0.27	0.22	-0.08	0.27	0.24	-0.14	0.27	-0.03	0.04
4. Army Career Intentions	0.51	0.50	0.25		0.36	0.35	0.18	-0.01	0.26	0.25	-0.17	0.31	-0.01	0.04
5. Army Reenlistment Intentions	0.51	0.51	0.26	0.85		0.87	0.23	-0.11	0.32	0.27	-0.25	0.28	-0.02	0.06
6. Army Life Adjustment	0.39	0.62	0.35	0.30	0.33		0.22	-0.11	0.31	0.27	-0.23	0.28	-0.02	0.08
7. Motivation to Lead: Affective	0.34	0.28	0.25	0.22	0.21	0.18		-0.18	0.63	0.50	-0.15	0.40	-0.06	0.15
8. Motivation to Lead: Noncalculative	-0.14	-0.38	-0.19	-0.14	-0.16	-0.43	-0.11		-0.30	-0.28	0.28	-0.18	0.07	-0.05
9. Motivation to Lead: Social-Normative	0.48	0.50	0.39	0.26	0.27	0.27	0.60	-0.30		0.55	-0.29	0.51	-0.04	0.12
10. Organizational Citizenship Behavior/Leadership	0.34	0.37	0.29	0.23	0.24	0.28	0.42	-0.29	0.50		-0.15	0.48	-0.13	0.15
11. Counterproductive Soldier Behavior	-0.20	-0.38	-0.23	-0.17	-0.17	-0.36	-0.12	0.37	-0.28	-0.16		-0.27	0.02	-0.05
12. Resilience	0.43	0.49	0.36	0.25	0.25	0.32	0.38	-0.21	0.55	0.46	-0.30		-0.13	0.15
13. Disciplinary Actions (Y/N)	-0.04	-0.06	-0.04	-0.05	-0.04	-0.14	-0.07	0.03	-0.04	-0.06	0.08	-0.08		-0.01
14. Last APFT Score	0.04	0.09	0.06	0.03	0.03	0.23	0.14	-0.06	0.10	0.12	-0.04	0.12	-0.12	

Note. Correlations below the diagonal are based on the full IMT criterion sample, $n = 57,794 - 139,403$. Correlations above the diagonal are based on the full in-unit criterion sample, $n = 350-16,653$. Correlations in bold are statistically significant ($p < .05$, two-tailed).

Table C.2. Correlations among the Peer Performance Rating Scales (PRS-P) in the Full IMT and In-Unit Criterion Samples

Score/Scale	1	2	3	4	5	6	7	8	9	10	11
1. Overall Performance Composite		0.85	0.81	0.71	0.80	0.84	0.84	0.35	0.80	0.82	0.88
2. Effort and Discipline	0.84		0.72	0.55	0.65	0.70	0.70	0.17	0.69	0.69	0.77
3. Working Effectively with Other Soldiers	0.83	0.75		0.51	0.61	0.67	0.68	0.13	0.64	0.65	0.73
4. Physical Fitness	0.74	0.53	0.50		0.51	0.54	0.58	0.12	0.52	0.55	0.61
5. MOS Qualification Knowledge & Skill	0.82	0.69	0.64	0.55		0.71	0.67	0.13	0.58	0.66	0.73
6. Processing Information & Solving Problems	0.86	0.72	0.70	0.61	0.70		0.73	0.12	0.65	0.71	0.76
7. Resilience and Adjustment								0.13	0.64	0.70	0.75
8. Counterproductive Soldier Behaviors	-0.32	-0.51	-0.47	-0.27	-0.39		-0.41		0.13	0.10	0.16
9. Going Above and Beyond	0.84	0.73	0.72	0.52	0.66		0.69	-0.42		0.68	0.70
10. Emergent Leadership											0.73
11. Overall Performance Rating	0.89	0.78	0.75	0.66	0.76		0.79	-0.47		0.75	

Note. Correlations below the diagonal reflect the IMT PRS-P, $n = 53,399$ -140,785. Correlations above the diagonal reflect the in-unit PRS-P, $n = 3,213$ -3,318. Missing values reflect the PRS-P that were not administered in either IMT. Ratings on PRS-P range from 1 and 5. PRS-P ratings from peers with a familiarity rating of 1 ("I have had little opportunity to observe this Soldier") were excluded from analyses. All correlations are statistically significant ($p < .01$, one-tailed).

Table C.3. Cross-instrument Correlations among Validation Analysis Criterion Scores in the Full IMT and In-Unit Criterion Samples

Criterion	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Knowledge & Skill ^a		--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2. WTBD JKT	0.80		0.13	0.05	0.08	-0.04	0.06	0.06	0.17	0.10	-0.01	-0.02	--	0.01	-0.03	-0.07
3. Commitment & Fit ^a	0.12	0.16		0.60	0.41	-0.28	0.46	0.08	0.14	0.12	-0.11	-0.02	0.01	-0.01	-0.05	-0.08
4. Retention Cognitions ^a	0.02	0.02	0.52		0.29	-0.25	0.29	0.08	0.06	0.11	-0.09	-0.01	--	--	-0.04	-0.08
5. Peer Leadership ^a	0.03	0.01	0.41	0.23		-0.11	0.48	0.15	0.13	0.14	-0.07	-0.03	--	--	-0.04	-0.08
6. CSB	-0.15	-0.13	-0.32	-0.18	-0.05		-0.27	-0.05	-0.05	-0.09	0.14	0.02	--	--	-0.02	0.02
7. Resilience	0.08	0.06	0.52	0.27	0.47	-0.30		0.15	0.16	0.14	-0.12	-0.03	--	--	-0.02	-0.05
8. APFT	0.04	0.06	0.07	0.03	0.12	-0.04	0.12		0.18	0.19	-0.11	-0.06	--	-0.02	-0.05	-0.11
9. PRS-S Overall Performance ^a	--	--	--	--	--	--	--	--		0.35	-0.18	-0.03	--	-0.05	-0.12	-0.18
10. PRS-P Overall Performance ^a	0.10	0.08	0.08	0.04	0.13	-0.08	0.14	0.27	--		-0.18	-0.04	--	--	-0.11	-0.16
11. Disciplinary Actions (Y/N)	-0.06	-0.03	-0.06	-0.05	-0.06	0.08	-0.08	-0.12	--	-0.17		0.06	--	--	0.04	0.05
12. IMT Restarts	-0.01	-0.01	-0.03	0.01	-0.02	0.02	-0.02	-0.04	--	-0.05	0.07		0.01	0.01	0.01	0.02
13. 6-mo Attrition	-0.01	-0.01	-0.05	-0.03	-0.02	0.01	-0.03	-0.05	--	-0.05	0.03	0.01		0.52	0.26	0.19
14. 12-mo Attrition	-0.04	-0.03	-0.07	-0.06	-0.03	0.03	-0.05	-0.06	--	-0.09	0.06	0.01	0.52		0.50	0.36
15. 24-mo Attrition	-0.08	-0.09	-0.08	-0.06	-0.04	0.04	-0.07	-0.11	--	-0.16	0.08	0.01	0.26	0.50		0.73
16. 36-mo Attrition	-0.13	-0.09	-0.05	-0.05	0.08	-0.11	-0.16	--	-0.25	0.15	0.02	0.19	0.36	0.73		

Note. ^aIndicates composite variable. CSB = Counterproductive Soldier Behavior. Correlations below the diagonal reflect the IMT criteria, $n = 8,174 - 139,403$. Correlations above the diagonal reflect the in-unit criteria, $n = 1,326 - 145,610$. Missing values reflect that the scales that were not administered at the relevant time point or that there was no attrition among those with data for the relevant criterion. Correlations in bold are statistically significant, $p < .05$ (two-tailed).

Table C.4. Group Differences on Key Criteria by Gender in the IMT and In-unit Validation Samples

Measure	Male		Female		<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
<i>IMT</i>					
Knowledge & Skill	0.06	0.86	-0.07	0.77	-0.15
Commitment & Fit	3.78	0.65	3.77	0.62	-0.01
Retention Cognitions	3.56	1.08	3.72	1.04	0.15
Peer Leadership	3.32	0.44	3.29	0.43	-0.07
PRS: Overall Performance - Peer	3.49	0.60	3.42	0.63	-0.11
WTBD JKT	0.58	0.15	0.55	0.16	-0.19
APFT	250.60	28.31	252.81	29.17	0.08
Counterproductive Soldier Behavior	1.96	0.63	1.81	0.57	-0.24
Resilience	4.01	0.61	4.00	0.57	-0.01
IMT Restarts	0.05	0.22	0.07	0.25	0.06
Disciplinary Incidents	0.23	0.42	0.23	0.42	0.01
<i>In Unit</i>					
Knowledge & Skill	-0.06	0.97	-0.31	0.95	-0.26
Commitment & Fit	3.32	0.79	3.25	0.72	-0.10
Retention Cognition	2.92	1.33	3.02	1.29	0.08
Peer Leadership	3.29	0.48	3.21	0.46	-0.18
PRS: Overall Performance - Supervisor	4.88	1.15	4.85	1.19	-0.02
PRS: Overall Performance - Peer	3.71	0.70	3.70	0.80	-0.02
WTBD JKT	0.59	0.14	0.55	0.14	-0.28
APFT	251.86	28.63	248.62	29.98	-0.11
Counterproductive Soldier Behavior	2.04	0.60	1.85	0.56	-0.32
Resilience	3.91	0.65	3.80	0.68	-0.17
PRS: Effort and Discipline	4.85	1.36	4.82	1.39	-0.02
PRS: Working with Others	4.96	1.24	5.03	1.27	0.06
PRS: Fitness and Bearing	4.91	1.56	4.70	1.65	-0.14
PRS: Emergent Leadership	4.46	1.66	4.49	1.67	0.02
Disciplinary Incidents	0.25	0.43	0.26	0.44	0.02
<i>Attrition</i>					
6-Month	0.08	0.28	0.13	0.34	0.16
12-Month	0.13	0.33	0.20	0.40	0.23
24-Month	0.21	0.40	0.31	0.46	0.26
36-Month	0.30	0.46	0.43	0.50	0.27

Note. Females are the referent group. Sample size for males ranged from 2,411 to 566,121; females ranged from 344 to 138,119. Significant Cohen's *d* values, based on an independent sample *t*-test between the group means are in bold ($p < .05$, two-tailed).

Table C.5. Group Differences on Key Criteria by Race and Ethnicity in the IMT and In-unit Validation Samples

Measure	White, Non-Hispanic		Black		<i>d</i>	Asian			Hispanic		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		<i>M</i>	<i>SD</i>	<i>d</i>	<i>M</i>	<i>SD</i>	<i>d</i>
<i>IMT</i>											
Knowledge & Skill	0.17	0.83	-0.35	0.86	-0.63	-0.02	0.81	-0.23	-0.16	0.82	-0.40
Commitment & Fit	3.80	0.65	3.70	0.66	-0.16	3.66	0.63	-0.21	3.80	0.63	-0.01
Retention Cognitions	3.53	1.09	3.72	1.07	0.18	3.61	1.02	0.08	3.65	1.05	0.11
Peer Leadership	3.30	0.42	3.36	0.48	0.14	3.29	0.46	-0.03	3.31	0.45	0.01
PRS: Overall Performance - Peer	3.46	0.61	3.47	0.63	0.03	3.52	0.59	0.10	3.52	0.58	0.11
WTBD JKT	0.60	0.15	0.51	0.16	-0.58	0.56	0.15	-0.23	0.55	0.15	-0.34
APFT	250.29	28.64	251.67	28.10	0.05	251.77	27.91	0.05	252.46	28.24	0.08
Counterproductive Soldier Behavior	1.95	0.61	1.96	0.70	0.01	1.90	0.60	-0.08	1.85	0.61	-0.17
Resilience	3.98	0.59	4.02	0.64	0.07	4.01	0.61	0.06	4.07	0.60	0.15
IMT Restarts	0.05	0.22	0.07	0.26	0.09	0.05	0.22	-0.00	0.05	0.22	0.01
Disciplinary Incidents	0.22	0.42	0.29	0.45	0.15	0.20	0.40	-0.06	0.22	0.41	-0.02
<i>In Unit</i>											
Knowledge & Skill	0.14	0.91	-0.48	1.00	-0.66	-0.22	0.91	-0.39	-0.22	0.94	-0.39
Commitment & Fit	3.33	0.82	3.24	0.73	-0.12	3.31	0.71	-0.03	3.3665	0.75	0.04
Retention Cognition	2.80	1.36	3.14	1.26	0.26	2.86	1.20	0.04	3.02	1.34	0.16
Peer Leadership	3.28	0.45	3.30	0.51	0.03	3.29	0.49	0.02	3.27	0.49	-0.03
PRS: Overall Performance - Supervisor	4.90	1.16	4.78	1.18	-0.10	4.98	1.10	0.07	4.89	1.14	-0.00
PRS: Overall Performance - Peer	3.67	0.72	3.72	0.70	0.06	3.77	0.66	0.14	3.78	0.72	0.15
WTBD JKT	0.61	0.13	0.53	0.14	-0.62	0.57	0.13	-0.33	0.56	0.14	-0.41
APFT	251.00	28.86	251.66	28.44	0.02	250.77	29.05	-0.01	252.72	28.92	0.06
Counterproductive Soldier Behavior	2.07	0.58	2.00	0.63	-0.11	1.95	0.61	-0.20	1.92	0.59	-0.25
Resilience	3.88	0.63	3.89	0.67	0.02	3.86	0.64	-0.03	3.95	0.67	0.11
PRS: Effort and Discipline	4.88	1.37	4.69	1.41	-0.14	5.00	1.24	0.09	4.87	1.32	-0.00
PRS: Working with Others	5.00	1.24	4.87	1.27	-0.10	5.03	1.21	0.02	4.93	1.23	-0.06
PRS: Fitness and Bearing	4.87	1.56	4.86	1.61	-0.01	5.01	1.54	0.09	4.93	1.55	0.04
PRS: Emergent Leadership	4.48	1.66	4.37	1.68	-0.07	4.45	1.64	-0.02	4.48	1.65	-0.01
Disciplinary Incidents	0.25	0.43	0.32	0.47	0.16	0.18	0.38	-0.16	0.23	0.42	-0.04
<i>Attrition</i>											
6-Month	0.10	0.31	0.08	0.28	-0.07	0.06	0.23	-0.16	0.07	0.26	-0.11
12-Month	0.16	0.36	0.13	0.33	-0.08	0.09	0.28	-0.20	0.10	0.31	-0.15
24-Month	0.25	0.43	0.22	0.42	-0.05	0.14	0.35	-0.24	0.17	0.38	-0.18
36-Month	0.35	0.48	0.34	0.47	-0.02	0.21	0.41	-0.29	0.26	0.44	-0.19

Note. The Minority group is the referent group. Sample size for White Non-Hispanic ranges from 1,339 to 397,552; Black from 634 to 154,203; Hispanic from 621 to 111,782; Asian from 142 to 31,912. Significant Cohen's *d* values, based on an independent sample *t*-test between the group means are in bold ($p < .05$, two-tailed).

APPENDIX D

BIVARIATE CORRELATIONS AMONG PREDICTOR AND CRITERION SCORES

Table D.1. Bivariate Correlations among AFQT, TAPAS, and IMT Criterion Scores

	Knowledge & Skill	WTBD JKT	Commitment & Fit	Retention Cognitions	Peer Leadership	CSB	Resilience
<i>n</i>	3,169 - 46,701	4,244 - 78,283	4,577 - 92,537	4,554 - 92,183	146 - 48,645	146 - 48,646	146 - 48,646
<i>AFQT</i>	.43	.35	-.04	-.11	-.04	.01	-.03
<i>Individual TAPAS Facets</i>							
Achievement	.04	.05	.13	.07	.16	-.12	.15
Adjustment	.08	.07	.03	.00	-.01	-.03	.04
Attention Seeking	-.01	.00	.04	.00	.13	.02	.08
Dominance	.04	.06	.10	.06	.28	-.05	.14
Even Tempered	.05	.06	.04	.02	-.01	-.09	.07
Intellectual Efficiency	.15	.13	.04	.02	.14	-.03	.08
Non-Delinquency	-.01	.00	.07	.03	.00	-.13	.05
Optimism	.04	.05	.08	.02	.05	-.07	.11
Order	-.09	-.08	.01	.03	.08	-.07	.06
Physical Conditioning	.00	.02	.05	.00	.11	.00	.14
Sociability	-.10	-.05	.08	.06	.18	-.02	.11
Team Orientation ^a	-.04	-.01	.07	.04	.06	-.04	.07
Tolerance	-.01	-.01	.03	.05	.03	-.07	.07
<i>TAPAS Composites</i>							
Can-Do	.21	.18	.01	-.02	.02	-.01	.01
Will-Do	.03	.05	.12	.04	.22	-.07	.20
Adaptation	.09	.09	.03	-.03	-.01	.00	.08

Note. WTBD JKT = Warrior Tasks and Battle Drills Job Knowledge Test. CSB = Counterproductive Soldier Behavior. PRS = Performance Rating Scales. Correlations are not reported where $n < 100$. Correlations in bold are statistically significant ($p < .05$, two-tailed).

^a Sample sizes for six TAPAS facets were considerably smaller than the other dimensions because not all facets were administered in every version of the TAPAS.

Table D.1. (Continued)

	PRS-P: Overall			
	APFT	Performance Composite	Disciplinary Incidents (Y/N)	Training Restarts
<i>n</i>	4,405 - 87,546	121 - 43,355	4,577 - 91,758	64,300 - 704,240
<i>AFQT</i>	.08	.06	-.01	.00
<i>Individual TAPAS Facets</i>				
Achievement	.11	.08	-.07	-.01
Adjustment	.01	-.09	-.02	.00
Attention Seeking	.09	.02	-.01	.00
Dominance	.12	.06	-.04	-.01
Even Tempered	-.02	.02	-.02	.01
Intellectual Efficiency	.04	.01	-.02	.01
Non-Delinquency	-.03	.01	-.03	.00
Optimism	.05	.04	-.04	-.01
Order	.04	.03	-.04	.00
Physical Conditioning	.26	.12	-.08	-.03
Sociability	.04	-.02	.00	.01
Team Orientation ^a	.03	.03	-.03	-.01
Tolerance	.00	-.01	.00	.02
<i>TAPAS Composites</i>				
Can-Do	.01	.03	-.01	.00
Will-Do	.26	.13	-.09	-.03
Adaptation	.18	.11	-.06	-.03

Note. WTBD JKT = Warrior Tasks and Battle Drills Job Knowledge Test. CSB = Counterproductive Soldier Behavior. PRS = Performance Rating Scales. Correlations are not reported where $n < 100$. Correlations in bold are statistically significant ($p < .05$, two-tailed).

^aSample sizes for six TAPAS facets were considerably smaller than the other dimensions because not all facets were administered in every version of the TAPAS.

Table D.2. Bivariate Correlations among AFQT, TAPAS and In-Unit Criterion Scores

	Knowledge & Skill	WTBD JKT	Commitment & Fit	Retention Cognitions	Peer Leadership	CSB	Resilience
<i>n</i>	862 - 6,473	862 - 6,467	1,021 - 9,001	168 - 3,867	168 - 3,869	168 - 3,869	168 - 3,869
<i>AFQT</i>	.43	.38	-.03	-.08	-.05	.08	-.03
<i>Individual TAPAS Facets</i>							
Achievement	.03	.05	.07	.01	.14	-.11	.12
Adjustment	.08	.07	.04	.05	.01	.03	.02
Attention Seeking	-.02	-.01	.00	.00	.09	.02	.05
Dominance	.02	.04	.04	.01	.22	-.02	.13
Even Tempered	.05	.07	.05	.03	-.02	-.10	.04
Intellectual Efficiency	.11	.11	.02	.00	.09	-.05	.07
Non-Delinquency	-.04	-.01	.06	.04	.00	-.09	.01
Optimism	.01	.00	.05	.01	.05	-.02	.13
Order	-.12	-.10	.00	.02	.08	-.09	.06
Physical Conditioning	.00	.02	.01	-.01	.10	.02	.12
Sociability	-.12	-.08	.06	.04	.14	-.03	.09
Team Orientation ^a	-.09	-.06	.03	.02	.09	-.02	.07
Tolerance	-.02	-.03	.03	-.01	.01	-.04	.04
<i>TAPAS Composites</i>							
Can-Do	.21	.20	-.01	-.02	.00	-.01	.01
Will-Do	.01	.04	.05	.00	.19	-.03	.18
Adaptation	.10	.10	.01	-.02	.01	.02	.07

Note. WTBD JKT = Warrior Tasks and Battle Drills Job Knowledge Test. CSB = Counterproductive Soldier Behavior. PRS = Performance Rating Scales. Correlations are not reported where $n < 100$. Correlations in bold are statistically significant ($p < .05$, two-tailed).

^a Sample sizes for six TAPAS facets were considerably smaller than the other dimensions because not all facets were administered in every version of the TAPAS.

Table D.2. (Continued)

	APFT	PRS-S: Overall Performance Composite	PRS-P: Overall Performance Composite	Disciplinary Incidents (Y/N)
<i>n</i>	990 - 8,634	723 - 6,265	169 - 2,755	1,038 - 9,163
<i>AFQT</i>	-.01	.09	.05	-.02
<i>Individual TAPAS Facets</i>				
Achievement	.08	.08	.08	-.07
Adjustment	.02	.02	-.12	-.01
Attention Seeking	.08	.02	-.02	.01
Dominance	.10	.06	.05	-.01
Even Tempered	-.03	.01	-.01	-.05
Intellectual Efficiency	.00	.04	.00	.00
Non-Delinquency	-.07	.00	-.01	-.06
Optimism	.01	.03	.05	-.04
Order	.03	.01	.06	-.05
Physical Conditioning	.22	.06	.06	-.04
Sociability	.05	.00	-.02	.02
Team Orientation ^a	.00	.00	.06	.00
Tolerance	-.04	-.02	-.02	.01
<i>TAPAS Composites</i>				
Can-Do	-.01	.04	.00	.00
Will-Do	.22	.09	.09	-.06
Adaptation	.13	.05	.05	-.05

Note. WTBD JKT = Warrior Tasks and Battle Drills Job Knowledge Test. CSB = Counterproductive Soldier Behavior. PRS = Performance Rating Scales. Correlations are not reported where $n < 100$. Correlations in bold are statistically significant ($p < .05$, two-tailed).

^aSample sizes for six TAPAS facets were considerably smaller than the other dimensions because not all facets were administered in every version of the TAPAS.

Table D.3. Bivariate Correlations among AFQT, TAPAS, and Attrition Criterion Scores (Regular Army Only)

	Attrition			
	6-Mos	12-Mos	24-Mos	36-Mos
<i>n</i>	56,239 - 629,138	45,759 - 507,130	41,207 - 427,117	39,475 - 364,488
<i>AFQT</i>	-.05	-.05	-.07	-.08
<i>Individual TAPAS Facets</i>				
Achievement	-.02	-.02	-.03	-.03
Adjustment	-.01	-.01	.00	.00
Attention Seeking	-.03	-.03	-.01	-.01
Dominance	-.03	-.03	-.02	-.02
Even Tempered	.00	.00	-.01	-.01
Intellectual Efficiency	.00	.00	.01	.01
Non-Delinquency	.02	.02	.00	.00
Optimism	-.03	-.03	-.03	-.04
Order	.01	.01	.01	.01
Physical Conditioning	-.08	-.09	-.09	-.09
Sociability	.01	.01	.02	.03
Team Orientation ^a	-.03	-.03	-.03	-.04
Tolerance	.01	.02	.02	.02
<i>TAPAS Composites</i>				
Can-Do	-.01	-.01	-.02	-.02
Will-Do	-.08	-.08	-.08	-.08
Adaptation	-.08	-.08	-.09	-.10

Note. Correlations in bold are statistically significant ($p < .05$, two-tailed).

^a Sample sizes for six TAPAS facets were considerably smaller than the other dimensions because not all facets were administered in every version of the TAPAS.