

Technical Report 1380

Tier One Performance Screen Initial Operational Test and Evaluation: Capstone Report

Deirdre J. Knapp, Editor

Human Resources Research Organization

Cristina D. Kirkendall, Editor

U.S. Army Research Institute



January 2020

**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:

**MICHELLE L. ZBYLUT, Ph.D.
Director**

Research accomplished under contract W911NF16C0056
for the Department of the Army by

Human Resources Research Organization

Technical Review by

Alexander P. Wind, U.S. Army Research Institute
Kristophor Canali, U.S. Army Research Institute
Richard Hoffman, U.S. Army Research Institute

NOTICES

DISTRIBUTION: This Technical Report has been submitted to the Defense Information Technical Center (DTIC). Address correspondence concerning reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: DAPE-ARI-ZXM, 6000 6th Street (Bldg 1464 / Mail Stop 2610), Fort Belvoir, Virginia 22060-5610.

FINAL DISPOSITION: Destroy this Technical Report when it is no longer needed. Do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE					
1. REPORT DATE (dd-mm-yy) January 2020		2. REPORT TYPE Final		3. DATES COVERED (from. . . to) August 2009 to December 2018	
4. TITLE AND SUBTITLE Tier One Performance Screen Initial Operational Test and Evaluation: Capstone Report				5a. CONTRACT OR GRANT NUMBER W911NF-16-C-0056	
				5b. PROGRAM ELEMENT NUMBER 633007	
6. EDITORS Deirdre J. Knapp and Cristina D. Kirkendall				5c. PROJECT NUMBER A792	
				5d. TASK NUMBER 329a	
				5e. WORK UNIT NUMBER	
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 6 th Street (Bldg 1464/Mail Stop 5610) Ft. Belvoir, VA 22060-5610				10. MONITOR ACRONYM ARI	
				11. MONITOR REPORT NUMBER Technical Report 1380	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution Statement A: Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES ARI Research POC: Dr. Cristina D. Kirkendall					
14. ABSTRACT (<i>Maximum 200 words</i>): 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 the 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 (ARI) investigations, the Army selected one promising measure, the Tailored Adaptive Personality Assessment System (TAPAS), for an initial operational test and evaluation (IOT&E), beginning administration to applicants in 2009. Criterion data were compiled at 6-month intervals from administrative records, schools for selected military occupational specialties (MOS), and Soldiers in units. This report is the 17th in a series of biannual evaluations of the TAPAS and serves as a final capstone report for what has been called the <i>TOPS IOT&E</i> research program. The cumulative results suggest that several TAPAS scales significantly predict a number of criteria of interest, indicating that the measure holds promise for both selection and assignment purposes.					
15. SUBJECT TERMS Behavioral and social science Personnel Manpower Selection and assignment					
SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT Unlimited Unclassified	20. NUMBER OF PAGES 121	21. RESPONSIBLE PERSON Tonia S. Heffner 703-545-4408
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified			

Technical Report 1380

**Tier One Performance Screen Initial Operational Test
and Evaluation: Capstone Report**

Deirdre J. Knapp, Editor

Human Resources Research Organization

Cristina D. Kirkendall, Editor

U.S. Army Research Institute

Selection and Assignment Research Unit

Tonia S. Heffner, Chief

January 2020

Approved for public release; distribution is unlimited

ACKNOWLEDGEMENTS

There are individuals not listed as authors who made significant contributions to the research described in this report. First and foremost are 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. Thanks also go to Dr. Tonia Heffner, Dr. Len White, and Ms. Sharon Meyers (ARI); Mr. Doug Brown, Mr. Chris Graves, Ms. Charlotte Campbell, and Mr. Bill Holden (HumRRO); and Ms. Winnie Young for their important contributions to this research effort.

We also extend our appreciation to the Army Test Program Advisory Team (ATPAT), a group of senior NCOs who periodically meet with 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. Over the course of this research, members of the ATPAT have included:

CW4 Jeremy Addleman	SGM Kenan Harrington	SSG Suhun Sung
SFC Eli Aguilar	SSG Teddy Hayes	SFC James Swenson
Mr. Liston Bailey	SFC William Hayes	SFC Steven Toslin
Ms. Kimberly Baker	1SG Thomas Hill Jr.	SGM Robert Trawick, Jr.
Mr. Shawn Bova	SFC Larry Howard Jr.	SGM Bert Vaughan
Mr. Tom Bowles	SGM Donnie Isaacs	MAJ Paul Walton
CSM Eddie Camp Jr.	SFC Vanda Johnson	MSG Clinton Washington
CSM John R. Calpena	SGM Thomas Klingel	Mr. Thomas Wetzel Sr.
MSG Patricio Cardonavega	Mr. James Lewis	SFC Kenneth Williams
SFC Timothy Castelow	SFC Ricardo Martinez	MSG Robert D. Wyatt
SSG William Cheek	Mr. Clifford McMillan	
CSM Lamont Christian	MSG Thomas Morgan	
MSG Fredrick Clayton	CW5 Anthony Moschella	
CSM Thomas Clingel	SGM Henry C. Myrick	
CSM Michael Cosper	CW4 Mary Negron	
SGM Daniel E. Dupont Sr.	CSM John Pack Jr.	
SGM John Eubank	Mr. William Palya	
MSG Christopher Evans	Mr. John Plotts	
CSM Michael Evans	Mr. Gerald Purcell	
MSG Tyler Ewing	MSG Joel Raglin	
1SG Robert Fortenberry	SGM Gregory A. Richardson	
SFC William Gokey	SGM Richard Rosen	
MSG Daniel Griego	CSM James Schultz	
Mr. Todd Grisso	SFC Michelle Schrader	
MSG Daniel Gutierrez	SGM Martez Sims	
SFC April Hansberry	SGM Brent Smith	
SFC Quinshaun R. Hawkins	SGM Christopher Smith	
CSM Brian A. Hamm	CSM (R) Clarence Stanley	
SFC Jerome Hanson	Mr. Robert Steen	

TIER ONE PERFORMANCE SCREEN INITIAL OPERATIONAL TEST AND EVALUATION: CAPSTONE REPORT

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 the 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)*. The TAPAS capitalizes on the latest advances in testing technology to assess motivation through the measurement of personality characteristics.

Procedure:

In May 2009, the U.S. Army began administering the TAPAS on the computer adaptive platform for the ASVAB (CAT-ASVAB) at Military Entrance Processing Stations (MEPS). To evaluate the TAPAS, outcome (criterion) data were collected at multiple points in time from Soldiers who took the TAPAS at entry. By the end of the IOT&E, initial military training (IMT) criterion data were being collected at schools for Soldiers in 71 military occupational specialties (MOS), a substantial increase over the original eight that were targeted at the beginning of the project. Project teams also collected criterion data from in-unit Soldiers (regardless of MOS) in multiple waves of site visits during the course of the IOT&E.

The criterion measures included job knowledge tests, an attitudinal assessment (the Army Life Questionnaire), and performance rating scales completed by the Soldiers' training cadre members (in IMT), supervisors (in units), and/or peers. To the extent possible, data from administrative records (e.g., separation status) were obtained for all participating Soldiers.

The data presented in this report come from TAPAS data collected through the spring of 2018 and criterion data collected through May 2018. In addition to reporting the most recent results, this report also stands as a capstone for the research conducted as part of the TOPS IOT&E since its inception in 2009. The foundational database consists of over a million (1,132,118) applicants who took the TAPAS; 1,048,245 of these individuals were in the TOPS Applicant Sample. The Applicant Sample excluded prior service applicants and those individuals ineligible for service based on education requirements or extremely low AFQT scores. The validation sample sizes were considerably smaller, with the IMT Validation Sample comprising 69,525 Soldiers, the In-

Unit Validation Sample comprising 8,513 Soldiers, and the Administrative Validation Sample (which includes Soldiers with criterion data [e.g., attrition] from at least one administrative source) comprising 621,590 Soldiers.

Data from the job knowledge tests, attitudinal assessment, performance ratings provided by cadre, 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 and theoretically consistent psychometric properties. The exception to this was the IMT cadre performance-rating scales and both in-unit and IMT peer rating scales, all of which exhibited low inter-rater reliability. Although we were unable to compute reliability estimates for the in-unit ratings, because there is only one supervisor rater per Soldier, it is likely that they are relatively unstable as well. As a result, findings involving the rating scales may underestimate relationships with other variables.

Multiple versions of the TAPAS have been administered at MEPS since its first use by the Army. As such, data from the TAPAS reflect not one but many versions of the measure, each with some number of different items. Similarly, data from each version were collected from different samples of Soldiers and, in some cases, at different periods of time. Given these factors, we conducted a meta-analysis to account for the different versions and to provide more accurate estimates of the relationships each TAPAS composite has with key Soldier outcomes.

Our approach to analyzing the TAPAS' incremental predictive validity over AFQT 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 the TAPAS composite or TAPAS scale scores were added to the baseline regression models served as our index of incremental validity. Correlations between TAPAS scale scores and selected criteria were also examined. Analyses used the TAPAS Will-Do, Can-Do, and Adaptation composite scores.

In prior evaluation reports, we conducted analyses that assessed implementation of the TAPAS with respect to AFQT categories and TAPAS score percentiles to further examine the relationships between the TAPAS and the various IMT, in-unit, and attrition criteria. For some of these analyses, AFQT Category IIIB/IV Soldiers were classified into two groups – those scoring below the 10th percentile on TAPAS and those scoring above the 10th percentile, to reflect prior Army TAPAS screening decisions.

Findings:

Results presented here align with the findings from prior analysis cycles. Overall, the TAPAS composites demonstrated incremental validity over the AFQT in predicting several important first-term Soldier outcomes. In particular, the TAPAS Will-Do composite demonstrated the greatest incremental validity overall. Of the eight Will Do criteria, ΔR estimates for six IMT and three in-unit criteria exceeded .05. The TAPAS Adaptation composite also yielded incremental validity

estimates above .05 for three IMT criteria and two in-unit criteria. In addition, both the Will-Do and Adaptation composites demonstrated negative relationships with the 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. Findings from the meta-analyses provided additional support for these observed relationships among the TAPAS composites and various outcomes. In general, the strongest meta-analytic corrected correlations were seen for outcomes where TAPAS provided the greatest prediction beyond the AFQT in the incremental validity analyses. Conversely, the TAPAS Can-Do composite did not provide any notable incremental validity for the criteria examined here. This finding is not surprising given the established strength of the AFQT in the prediction of cognitively-based criteria. Nonetheless, the Can-Do composite had small positive meta-analytic correlations with the cognitive outcomes.

As demonstrated in the last evaluation report (Knapp & Kirkendall, 2018), a distinction was also seen when comparing AFQT Category IIIB and IV Soldiers who scored above and below the 10th percentile on TAPAS. One particularly large difference was for disciplinary incidents where the low scoring groups had approximately 5% and 8% more disciplinary incidents, respectively, compared to the higher scoring IIIB and IV groups. The higher scoring Category IIIB and IV groups had lower attrition rates (ranging from a difference of 1.2% to 5.6%) than the low scoring groups at 6, 12, and 24 months' time in service. The Adaptation composite generally provided small incremental validity gains for predicting attrition, showing relatively larger gains for predicting attrition later in the enlistment term. Even these small gains in validity are important, given the modest relationship with the AFQT and the cost for training each Soldier.

Utilization and Dissemination of Findings:

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

TIER ONE PERFORMANCE SCREEN INITIAL OPERATIONAL TEST AND EVALUATION: CAPSTONE REPORT

CONTENTS

	Page
CHAPTER 1: INTRODUCTION.....	1
Cristina D. Kirkendall (ARI), Deirdre J. Knapp (HumRRO), and Tonia S. Heffner (ARI)	
Background	1
The Tier One Performance Screen (TOPS).....	2
Evaluating TAPAS	3
Overview of Report	4
CHAPTER 2: SAMPLE CHARACTERISTICS	5
D. Matthew Trippe, Erin S. Banjanovic, and Justin Purl (HumRRO)	
Data Sources	5
Sample Filters	6
Description of Analysis Samples	7
Summary	10
CHAPTER 3: THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM (TAPAS)	11
Stephen Stark, O. Sasha Chernyshenko, Christopher Nye, and Fritz Drasgow (Drasgow Consulting Group)	
Description	11
Psychometric Properties of TAPAS Test Versions	14
TAPAS Composites	14
Summary	15
CHAPTER 4: DESCRIPTION AND PSYCHOMETRIC PROPERTIES OF OBJECTIVE CRITERION MEASURES	16
Christopher Graves and William Taylor (HumRRO)	
Job Knowledge Tests (JKTs)	16
Administrative Criteria.....	20
Summary	22
CHAPTER 5: ARMY LIFE QUESTIONNAIRE	23
Colby K. Nesbitt (HumRRO), Elizabeth Salmon (ARI), and Cristina D. Kirkendall (ARI)	
Background	23
Self-Report Administrative Criteria	23
Revisions to ALQ Attitudinal Scales	24
Development of New ALQ Content.....	26
ALQ Composite Scores.....	30
Final ALQ Contents and Associated Psychometric Properties	34
Summary	35

CONTENTS (CONTINUED)

Page

List of Tables

Table 2.1. Full TAPAS Sample Characteristics.....	8
Table 2.2. Background and Demographic Characteristics of the TOPS Samples	9
Table 3.1. TAPAS Facets Names and Definitions.....	12
Table 3.2. TAPAS Versions by Administration Date.....	14
Table 4.1. Job Knowledge Test (JKT) Reliability Estimates in the IMT and In-Unit Validation Samples	19
Table 4.2. Job Knowledge Test (JKT) Descriptive Statistics in the IMT Validation Sample	20
Table 4.3. Base Rates for Dichotomous Criteria in the Administrative Validation Sample	21
Table 4.4. Base Rates for Advanced Individual Training Grades in the Validation Sample	21
Table 5.1. Background and Demographic Characteristics of Pilot Sample.....	24
Table 5.2. Reliability Estimates of ALQ Scales, Before and After Revision.....	25
Table 5.3. Descriptive Statistics and Correlations between Scores on Existing ALQ Scales, Before and After Revision	25
Table 5.4. New Target Content for the Revised ALQ	26
Table 5.5. ALQ Composites and Associated Scales for the Original and Revised Versions (IMT and In-Unit).....	31
Table 5.6. Differences in Correlations for the Original and Revised ALQ Composites	32
Table 5.7. CFA Model Fit Statistics for the Army Life Questionnaire (ALQ) New and Revised Criterion Composites	34
Table 5.8. ALQ Scale Reliability Estimates	34
Table 5.9 Revised ALQ Descriptive Statistics	35
Table 6.1. IMT PRS-S Army-Wide Rating Scale Dimensions and Associated Definitions	39
Table 6.2. IMT PRS-S Interrater Reliability Estimates	40
Table 6.3. Descriptive Statistics for the Supervisor Performance Rating Scales (PRS-S) in the IMT Validation Sample	41
Table 6.4. In-Unit PRS-S Army-Wide Rating Scale Dimensions and Associated Definitions	42
Table 6.5. Descriptive Statistics for the Supervisor Army-Wide Performance Rating Scales (PRS-S) in the In-Unit Validation Sample	43
Table 6.6. IMT PRS-P Rating Scale Dimensions and Associated Example Behaviors	45
Table 6.7. In-Unit PRS-P Rating Scale Dimensions and Associated Example Behaviors.....	47
Table 6.8. Percentage of Soldiers with Various Numbers of Peer Raters	48
Table 6.9. PRS-P Interrater Reliability Estimates	49

CONTENTS (CONTINUED)

Page

List of Tables

Table 6.10. CFA Model Fit Statistics for the Peer Performance Rating Scales (PRS-P)	50
Table 6.11. Descriptive Statistics for the Peer Performance Rating Scales (PRS-P) in the IMT Validation Sample	51
Table 6.12. Descriptive Statistics for the Army-Wide Performance Rating Scales (PRS-P) in the In-Unit Validation Sample.....	51
Table 7.1. Criteria Used to Examine the Validity of the TAPAS.....	56
Table 7.2. Meta-analytic Results for TAPAS Correlations with IMT Technical Performance	60
Table 7.3. Meta-analytic Results for TAPAS Correlations with IMT ALQ and APFT Criteria	61
Table 7.4. Meta-analytic Results for TAPAS Correlations with IMT Performance Rating Criteria	61
Table 7.5. Meta-analytic Results for TAPAS Correlations with In-Unit Technical Performance	62
Table 7.6. Meta-analytic Results for TAPAS Correlations with In-Unit ALQ and APFT Criteria	63
Table 7.7. Meta-analytic Results for TAPAS Correlations with In-Unit Performance Rating Criteria.....	63
Table 7.8. Meta-analytic Results for TAPAS Correlations with Administrative Criteria	64
Table 7.9. Meta-analytic Results for TAPAS Correlations with Cumulative Attrition through 24 Months of Service (Regular Army Only).....	65
Table 7.10. Incremental Validity Estimates for the TAPAS over AFQT for Predicting IMT Technical Performance	66
Table 7.11. Incremental Validity Estimates for the TAPAS over AFQT for Predicting IMT ALQ and APFT Criteria.....	67
Table 7.12. Incremental Validity Estimates for the TAPAS over AFQT for Predicting IMT Performance Rating Criteria	67
Table 7.13. Incremental Validity Estimates for the TAPAS over AFQT for Predicting In- Unit Technical Performance Criteria	69
Table 7.14. Incremental Validity Estimates for the TAPAS over AFQT for Predicting In- Unit ALQ and APFT Criteria	70
Table 7.15. Incremental Validity Estimates for the TAPAS over AFQT for Predicting In- Unit Performance Rating Criteria	71
Table 7.16. Incremental Validity Estimates for the TAPAS Composites over AFQT for Predicting Dichotomous Criteria	73
Table 7.17. Incremental Validity Estimates for the TAPAS Composite Scores over AFQT for Predicting Cumulative Attrition through 24 Months of Service (Regular Army Only).....	74

CONTENTS (CONTINUED)

Page

List of Tables

Table B.1. Means and Standard Deviations for the TAPAS Scales and Composites.....	B-1
Table B.2. Correlations among the AFQT and TAPAS Scales and Composites in the Applicant Sample.....	B-2
Table B.3. Group Differences on the AFQT and TAPAS by Gender	B-3
Table B.4. Group Differences on the AFQT and TAPAS by Race and Ethnicity.....	B-4
Table C.1. Correlations among the Army Life Questionnaire (ALQ) Scales in the IMT and In-Unit Validation Samples	C-1
Table C.2. Correlations among the Supervisor Performance Rating Scales (PRS-S) in the IMT Validation Sample	C-3
Table C.3. Correlations among the Supervisor Performance Rating Scales (PRS-S) in the In-Unit Validation Sample.....	C-4
Table C.4. Correlations among the Peer Performance Rating Scales (PRS-P) in the IMT and In-Unit Validation Sample	C-5
Table C.5. Cross-instrument Correlations among Validation Analysis Criterion Scores in the IMT and In-Unit Validation Samples	C-6
Table E.1. Summary of the Bivariate Correlations between AFQT, TAPAS, and Selected IMT Criteria	E-1
Table E.2. Summary of the Bivariate Correlations between AFQT, TAPAS, and Selected In-Unit Criteria.....	E-3
Table E.3. Summary of the Bivariate Correlations between AFQT, TAPAS, and Selected Attrition Criteria (Regular Army Only).....	E-5

CONTENTS (CONTINUED)

Page

List of Figures

Figure 2.1. Overview of TOPS data file merging and nested sample generation process.....	5
Figure 6.1. Sample IMT PRS-S rating scale.....	39
Figure 6.2. Sample in-unit supervisor rater performance rating scale.....	43
Figure 6.3. Sample IMT peer rater performance rating scales.	46
Figure 7.1. Increase in prediction of IMT criteria using Will-Do TAPAS scales.	68
Figure 7.2. Increase in prediction of IMT criteria using Can-Do TAPAS scales.....	68
Figure 7.3. Increase in prediction of in-unit criteria using Will-Do TAPAS scales.....	71
Figure 7.4. Increase in prediction of in-unit criteria using Can-Do TAPAS scales.....	71
Figure 7.5. Predictive accuracy of the AFQT and Will-Do TAPAS composite in the discrimination of both IMT and In-Unit Disciplinary Incidents and IMT Restarts.....	76
Figure 7.6. Predictive accuracy of the AFQT and Adaptation TAPAS composite in the discrimination of attrition outcomes for Regular Army Soldiers.	77

TIER ONE PERFORMANCE SCREEN INITIAL OPERATIONAL TEST AND EVALUATION: CAPSTONE REPORT

CHAPTER 1: INTRODUCTION

Cristina D. Kirkendall (ARI), Deirdre J. Knapp (HumRRO), and Tonia S. Heffner (ARI)

Background

The Selection and Assignment Research Unit (SARU) of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is responsible for conducting personnel research for the Army. The focus of SARU's research is maximizing the potential of the individual Soldier through effective selection, assignment, and retention strategies.

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.¹ Although the 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).

In December 2006, the Department of Defense (DoD) ASVAB review panel—a panel of experts in the measurement of human characteristics and performance—released their recommendations (Dragow, Embretson, Kyllonen, & Schmitt, 2006), several of which focused on supplementing the ASVAB with additional measures for use in selection and assignment decisions. The ASVAB review panel further recommended that the use of these measures be validated against performance criteria.

Just prior to the release of the ASVAB review panel's findings, ARI had initiated a longitudinal research effort, *Validating Future Force Performance Measures (Army Class)*, to examine the prediction potential of several non-cognitive measures (e.g., temperament and person-environment fit) for 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, Knapp, & Owens, 2013; Ingerick, Diaz, & Putka, 2009; Knapp & Heffner, 2009). Experimental predictors were administered to new Soldiers in 2007 and early 2008. Army Class collected school-based criterion data on a subset of the Soldier sample as they completed job training. Job performance criterion data were collected from Soldiers in the Army Class longitudinal validation sample in 2009 with a second round of data collections in Soldiers' units completed in April 2011 (Knapp, Owens, & Allen, 2012). Final analysis and reporting of this program of research can be found in Allen et al. (2013).

¹ A list of defined acronyms commonly used in this report is provided in Appendix A.

After the Army Class research began, ARI initiated the *Expanded Enlistment Eligibility Metrics (EEEM)* project (Knapp & Heffner, 2010). The EEEM goals were similar to Army Class, but the focus was specifically on Soldier selection and the time horizon was much shorter. Specifically, EEEM required identification of one or more promising new predictor measures for immediate implementation. The EEEM project capitalized on the existing Army Class data collection procedure and, thus, the EEEM sample was a subset of the Army Class sample. Based on the EEEM findings, Army policy-makers approved an initial operational test and evaluation (IOT&E) of the *Tier One Performance Screen (TOPS)*.

The Tier One Performance Screen (TOPS)

Six experimental pre-enlistment measures were included in the EEEM research (Allen, Cheng, Putka, Hunter, & White, 2010). These included several temperament measures, a situational judgment test, and two person-environment fit measures based on values and interests. The most promising measures recommended to the Army for implementation were identified based on the following considerations:

- Incremental validity over AFQT for predicting important performance and retention-related outcomes,
- Minimal subgroup differences,
- Low susceptibility to response distortion (e.g., faking optimal responses), and
- Minimal administration time requirements.

The Tailored Adaptive Personality Assessment System ([TAPAS]; Stark, Chernyshenko, & Drasgow, 2010) surfaced as the top choice². The TAPAS is a measure of personality characteristics (e.g., achievement, sociability) that capitalizes on the latest advances in psychometric theory and provides a good indicator of personal motivation.

In May 2009, the Army began administering the TAPAS at Military Entrance Processing Stations (MEPS) on the computer adaptive platform for the ASVAB (CAT-ASVAB). Initially, the TAPAS was to be administered only to Education Tier 1, non-prior service applicants.³ This limitation to Education Tier 1 was removed early in 2011 to allow the Army to evaluate the TAPAS across all types of applicants.

The TOPS research investigated the potential for non-cognitive measures to identify 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 the TAPAS as a tool for “screening in” Education Tier 1 applicants with lower AFQT scores, allowing more of these applicants to access into the Army, hence the research program title (Tier One Performance Screen). However, changing economic conditions spurred a reconceptualization that led to using the TAPAS as a tool to screen out low motivated applicants, thus making the selection criteria to

² Other promising assessments include the Work Preferences Assessment ([WPA]; Putka & Van Iddekinge, 2007) and the Information/Communications Technology Literacy test ([ICTL]; Russell & Sellman, 2009).

³ Applicant educational credentials are classified as Tier 1 (primarily high school diploma), Tier 2 (primarily non-diploma graduate), and Tier 3 (not a high school graduate).

enter the Army more stringent. Decision rules regarding the derivation and use of TAPAS composite scores have been adjusted in myriad ways to reflect evolving market conditions.

In practice, as part of the TOPS IOT&E TAPAS scores were originally used to screen out a small number of AFQT Category IIIB/IV applicants.⁴ In 2014, Tier 2 applicants who scored below the 30th percentile on TAPAS were also screened out. In 2015, the IOT&E was modified to allow all Tier 1 Category IIIB applicants to enlist regardless of TAPAS score. Most recently (at the beginning of 2017), the Army temporarily suspended use of TAPAS for operational applicant screening, though TAPAS continues to be administered to most Army applicants.

Evaluating TAPAS

To evaluate the TAPAS, the Army collected training criterion data on Soldiers in multiple military occupational specialties (MOS) as they completed initial military training (IMT). This originally involved a target group of eight MOS which comprised large, highly critical MOS as well as MOS to represent the diversity of job requirements across the Army. Over time, however, additional MOS were added and the most recent target group included 71 MOS. The criterion measures included job knowledge tests (JKTs); an attitudinal assessment, the Army Life Questionnaire (ALQ); and performance rating scales (PRS) for cadre/instructors and peers. These measures were computer-administered at the end of Initial Military Training (IMT) for each of the target MOS. The process was overseen by Army personnel with guidance and support from both ARI and HumRRO. Administrative records were obtained for all Soldiers who took the TAPAS, regardless of MOS. Although data were collected from a large number of MOS, in most cases the sample sizes were too small for MOS-specific analysis.

Criterion data were also collected from Soldiers and their supervisors during data collection trips to major Army installations throughout the course of the IOT&E. These proctored “in-unit” data collections began in January 2011 and targeted all Soldiers who took the TAPAS prior to enlistment. The in-unit criterion measures included JKTs, the ALQ, supervisor ratings of performance, and peer ratings of performance. Separation status of all Soldiers who took the TAPAS prior to enlistment was tracked throughout the course of the research.

This report describes the 17th iteration of the criterion-related validation through the TOPS IOT&E initiative and serves as a capstone for the TOPS IOT&E. Prior evaluations are described in a series of technical reports (Bynum & Mullins, 2017; Knapp & Heffner, 2011, 2012; Knapp, Heffner, & White, 2011; Knapp & LaPort, 2013a, 2013b, 2014; Knapp & Kirkendall, 2018; Knapp & Wolters, 2017) and internal memoranda. Though this is the final report of the TOPS effort, this line of research will continue as part of a follow-on ARI research program entitled *Validation of Accessions Screening Tools (VAST)*.

⁴ Examinees are classified into categories based on their AFQT percentile scores (Category I = 93-99, Category II = 65-92, Category IIIA = 50-64, Category IIIB = 31-49, Category IV = 10-30, Category V = 1-9).

Overview of Report

Chapter 2 explains how the evaluation analysis data files have been constructed and then describes characteristics of multiple TOPS analysis samples. Chapter 3 describes the TAPAS, including content, scoring, and psychometric characteristics. Chapters 4 through 6 describe the IMT and in-unit criterion measures used in this evaluation, including how they have evolved since the start of the TOPS IOT&E program. Chapter 7 summarizes the criterion scores and describes the method and results of the criterion-related validation analyses. The report concludes with Chapter 8, which summarizes the TOPS IOT&E and looks forward to a program for the continuing validation of Army enlistment screening tools.

CHAPTER 2: SAMPLE CHARACTERISTICS

D. Matthew Trippe, Erin S. Banjanovic, and Justin Purl (HumRRO)

This chapter describes characteristics of the samples used in the TOPS IOT&E evaluation analyses. We begin with a summary of data sources, describe how Soldier data were filtered for analysis, and then describe multiple subsamples that were created to support various types of analyses.

Data Sources

An illustrative view of the TOPS sources of predictor and criteria data is provided in Figure 2.1 (see Appendix A for acronym definitions). The lighter boxes within the figure represent sources of data, and the darker boxes represent samples on which descriptive or inferential analyses were conducted. The leftmost column in the figure summarizes the predictor data sources used to derive the TOPS Applicant Sample. The other columns summarize the research-only (i.e., non-administrative) and administrative criterion data. The box shown with dashes reflects administrative data that have been unavailable to us for analysis since 2013. Predictor and criterion data were merged to form the IMT, In-Unit, and Administrative Validation Samples.

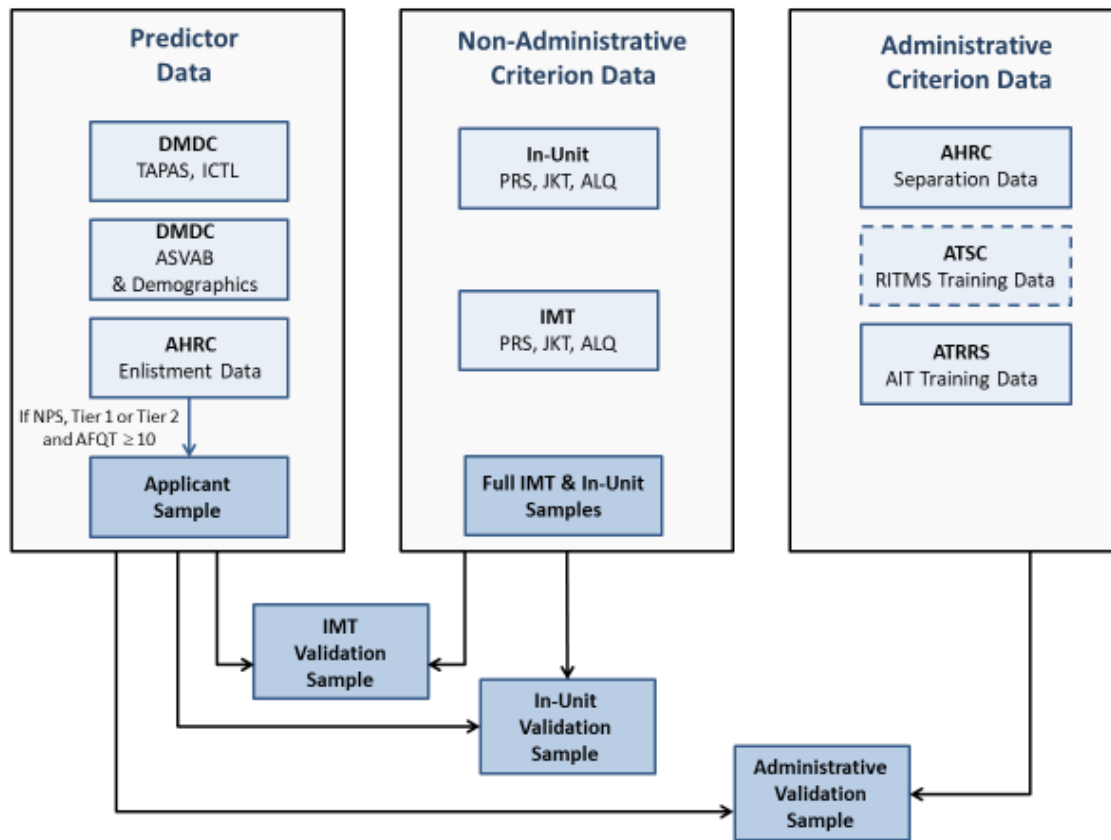


Figure 2.1. Overview of TOPS data file merging and nested sample generation process.

Sample Filters

For the purpose of evaluating the effectiveness of TAPAS, exclusions to the analysis samples were imposed based on AFQT score, education level, service history, component, and MOS.

AFQT Category

The ASVAB is a multiple aptitude battery of tests administered by the MEPCOM. Most military applicants take the computer adaptive version of the ASVAB (CAT-ASVAB). Scores on the ASVAB tests are combined to create composite scores for use in selecting applicants into the Army and qualifying them for MOS. The AFQT, the composite used for selecting applicants into the Army, comprises the Verbal Expression⁵ (VE), Arithmetic Reasoning (AR), and Math Knowledge (MK) tests ($AFQT = 2*VE + AR + MK$). Applicants must meet a minimum AFQT score to be eligible to serve in the military, and the Services favor high-scoring applicants for enlistment. AFQT percentile scores are divided into the following categories:⁶

- 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 DoD regulations, AFQT Category V applicants are not eligible for enlistment. Category IV accessions are greatly restricted, some restriction is placed on accessing Category IIIB accessions, and priority is given to Category I-III A accessions. The Applicant Sample excludes Soldiers with an AFQT score of less than 10 (i.e., Category V; $n = 23,061$).

AFQT descriptive statistics and subgroup score comparison information is reported in Appendix B. AFQT category frequencies are reported in Tables 2.1 and 2.2.

Education Tier

In the early 1980s, DoD initiated a detailed study of the relationship between educational credentials, other background characteristics, and adaptability for military service. The results supported a three-tier classification of educational credentials. The tiers have evolved over time and are now defined as:

- Tier 1 – Primarily high school diploma graduate and higher (e.g., individuals currently in high school or college, college graduates, adult/alternative diplomas, home school diplomas)
- Tier 2 – Primarily non-diploma graduate (e.g., GED certificants, vocational-technical certificants)

⁵ Verbal Expression is a scaled combination of the Word Knowledge (WK) and Paragraph Comprehension (PC) tests.

⁶ For more information on ASVAB scoring, see the official website of the ASVAB, www.officialasvab.com.

Tier 3 – Non-high school graduate (i.e., individuals not currently attending high school and do not possess a high school diploma or alternate credential)

Consistent with DoD policy, which specifies that Soldiers classified as Tier 3 are ineligible for accession, the Applicant Sample excludes Tier 3 Soldiers ($n = 7,224$) and those with unknown values ($n = 7,369$).

Service History

Because the TOPS program was designed to predict first term Soldier performance, individuals with prior service history are excluded from the analysis samples ($n = 18,149$).

Service Component

The Applicant Sample includes Soldiers from all Army components – Regular Army (RA), U.S. Army Reserve (USAR), and U.S. Army National Guard (ARNG). For most analyses, Soldiers from all components are included. However, for analyses involving separation data, only Regular Army Soldiers were included.

Description of Analysis Samples

Table 2.1 summarizes the full TAPAS sample by the key variables that were used to create the analysis samples. Among the 1,132,118 applicants in the total unfiltered sample, 1,048,245 (92.6%) met the screening criteria for the Applicant Sample. A detailed breakout of background and demographic characteristics observed in the analysis samples appears in Table 2.2.

The Administrative Validation Sample includes 621,590 Soldiers who meet all of the inclusion criteria for the TOPS Applicant Sample and also have at least one record in an administrative criterion data source (e.g., Army Training Requirements and Resources System [ATRRS], Resident Individual Training Management System [RITMS]). There are 109,498 Soldiers with IMT criterion data; however, only 69,525 were linked to an administrative TAPAS record and included in the IMT Validation Sample. Similarly, there are 14,557 Soldiers with in-unit data but only 8,513 of these Soldiers have matching TAPAS data and were included in the In-Unit Validation Sample. There are 1,044 Soldiers with a TAPAS record and both IMT and in-unit criterion data.

There are two primary reasons for the diminution of sample sizes between the Applicant Sample and the Administrative Validation samples. First, is the fact that many of the applicants did not access into the Army. Second, we relied on self-reported name and date of birth to match TAPAS records to the criterion data, which often resulted in unsuccessful matches. Further, fewer than half of the total number of Soldiers for whom we had IMT and in-unit criterion data were in the IMT and In-Unit Validation samples. In addition to cases lost due to unreliable reporting of the matching variables (name and date of birth), criterion testing started early in 2009 before TAPAS was being widely administered to applicants.

Table 2.1. Full TAPAS Sample Characteristics

Variables	<i>n</i>	% of Total Sample (<i>N</i> = 1,132,118)
<i>Education Tier</i>		
Tier 1	1,062,482	93.8
Tier 2	55,043	4.9
Tier 3	7,224	0.6
Unknown	7,369	0.7
<i>Prior Service</i>		
Yes	18,149	1.6
No	1,113,969	98.4
<i>Military Occupational Specialty</i>		
Infantryman (11B/C/X + 18X)	104,968	9.3
Combat Engineer (12B)	19,226	1.7
Cannon Crewmember (13B)	12,954	1.1
Field Artillery Automated Tactical Data Systems Specialist (13D)	4,807	0.4
Fire Support Specialist (13F)	7,460	0.7
Cavalry Scout (19D)	17,096	1.5
M1 Armor Crewman (19K)	7,840	0.7
Military Police (31B)	29,407	2.6
Human Resources Specialist (42A)	17,865	1.6
Combat Medic Specialist (68W)	32,211	2.8
Motor Transport Operator (88M)	33,798	3.0
Wheeled Vehicle Mechanic (91B)	35,493	3.1
Other	369,534	32.6
Unknown ^a	439,459	38.8
<i>AFQT Category</i> ^b		
I	62,218	5.5
II	308,214	27.2
IIIA	220,472	19.5
IIIB	349,058	30.8
IV	164,557	14.5
V	23,061	2.0
Unknown ^a	4,538	0.4
<i>Contract Status</i>		
Signed	741,752	65.5
Not signed	390,366	34.5
Applicant Sample ^c	1,048,245	92.6

^a Generally, when the MOS or AFQT category is unknown, it is either because the information was not yet available in the data sources on which the June 2018 data file was based or because the respondent did not access into the Army.

^b AFQT Categories IIIB and IV are oversampled. Values presented are not representative of Army accessions.

^c The Applicant Sample size is smaller than the total TAPAS sample because it is limited to MOS other than 09L and 09C, non-prior service, Education Tier 1 and 2, and AFQT ≥ 10 applicants.

Table 2.2. Background and Demographic Characteristics of the TOPS Samples

Characteristic	Applicant ^a <i>n</i> = 1,048,245		Administrative Validation ^b <i>n</i> = 621,590		IMT Validation ^c <i>n</i> = 69,525		In-Unit Validation ^d <i>n</i> = 8,513	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
<i>Component</i>								
Regular	597,297	57.0	364,524	58.6	42,577	61.2	8,450	99.3
ARNG	312,940	29.9	177,932	28.6	20,253	29.1	35 ^e	0.4
USAR	138,008	13.2	79,134	12.7	6,695	9.6	28 ^e	0.3
<i>Education Tier</i>								
Tier 1	996,074	95.0	596,265	95.9	66,899	96.2	8,252	96.9
Tier 2	52,171	5.0	25,325	4.1	2,626	3.8	261	3.1
<i>Military Occupational Specialty</i>								
Infantryman (11B/C/X + 18X)	99,056	9.4	96,161	15.5	17,515	25.2	2,027	23.8
Combat Engineer (12B)	18,493	1.8	16,827	2.7	4,899	7.0	325	3.8
Cannon Crewmember (13B)	12,359	1.2	11,576	1.9	250	0.4	208	2.4
Field Artillery Tactical Data Systems Specialist (13D)	4,580	0.4	4,561	0.7	962	1.4	96	1.1
Fire Support Specialist (13F)	7,069	0.7	6,824	1.1	1,127	1.6	102	1.2
Cavalry Scout (19D)	16,191	1.5	15,612	2.5	2,089	3.0	426	5.0
Armor Crewman (19K)	7,520	0.7	7,475	1.2	2,156	3.1	195	2.3
Military Police (31B)	27,907	2.7	26,102	4.2	8,552	12.3	168	2.0
Human Resources Specialist (42A)	16,935	1.6	16,239	2.6	2,256	3.2	225	2.6
Combat Medic Specialist (68W)	30,786	2.9	29,117	4.7	8,953	12.9	343	4.0
Motor Transport Operator (88M)	32,173	3.1	29,743	4.8	9,255	13.3	508	6.0
Wheeled Vehicle Mechanic (91B)	34,021	3.2	31,925	5.1	2,785	4.0	460	5.4
Other	349,294	33.3	329,313	53.0	8,725	12.5	3,430	40.3
Unknown	391,861	37.4	115	0.0	1	0.0	--	--
<i>AFQT Category^f</i>								
I	57,515	5.5	35,912	5.8	4,074	5.9	383	4.5
II	290,465	27.7	189,236	30.4	23,846	34.3	2,234	26.2
IIIA	209,248	20.0	135,006	21.7	14,647	21.1	1,857	21.8
IIIB	333,768	31.8	222,928	35.9	22,675	32.6	3,496	41.1
IV	157,249	15.0	38,508	6.2	4,283	6.2	543	6.4
<i>Gender</i>								
Female	229,750	21.9	114,797	18.5	9,636	13.9	1,066	12.5
Male	794,406	75.8	487,440	78.4	58,706	84.4	7,122	83.7
Missing	24,089	2.3	19,353	3.1	1,183	1.7	325	3.8
<i>Race</i>								
African American	248,475	23.7	136,848	22.0	11,473	16.5	1,986	23.3
American Indian	8,450	0.8	4,917	0.8	556	0.8	57	0.7
Asian	49,912	4.8	27,807	4.5	2,667	3.8	501	5.9
Hawaiian/Pacific Islander	2,515	0.2	1,327	0.2	149	0.2	28	0.3
Caucasian	721,497	68.8	443,371	71.3	53,685	77.2	5,796	68.1
Multiple	3,500	0.3	2,172	0.3	204	0.3	35	0.4
Declined to Answer/Missing	13,896	1.3	5,148	0.8	791	1.1	110	1.3

Table 2.2. (Continued)

Characteristic	Applicant ^a		Administrative Validation ^b		IMT Validation ^c		In-Unit Validation ^d	
	<i>n</i> = 1,048,245		<i>n</i> = 621,590		<i>n</i> = 69,525		<i>n</i> = 8,513	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
<i>Ethnicity</i>								
Hispanic/Latino	170,96	16.3	98,844	15.9	11,440	16.5	1,459	17.1
Not Hispanic	863,85	82.4	518,293	83.4	57,425	82.6	6,955	81.7
Declined to Answer/Missing	13,434	1.3	4,453	0.7	660	0.9	99	1.2

^a Limited to applicants who had no prior service, Education Tier 1 or 2, AFQT ≥ 10 ; served as the core analysis sample.

Additionally, 09L and 09C Soldiers were removed from all analyses.

^b Soldiers in Applicant Sample with at least one criterion record (i.e., schoolhouse, in-unit, ATRRS, RITMS, or attrition).

^c Soldiers in Applicant Sample with criterion data collected at schoolhouses.

^d Soldiers in Applicant Sample with criterion data collected in units.

^e We believe these Soldiers were on active duty when the in-unit data collections were taking place.

^f AFQT Categories IIIB and IV are oversampled. Values presented are not representative of Army accessions.

Sample sizes reported in all subsequent chapters and appendices are generally smaller than the figures reported here because of further data filtering or disaggregation that occurred. For example, predictor and criterion scores were determined to be valid if they passed multiple data quality screens intended to identify unmotivated responding. Additional screens were analysis specific and thus not applied to the descriptive analysis of the samples described in this chapter. Further, a small number of Soldiers in the Applicant Sample ($n = 1,534$) were administered an early version of the TAPAS (13D-CAT) and were excluded from analyses because of minor dissimilarities with subsequent TAPAS forms.

Summary

The TOPS analysis samples represent a combination of administrative, IMT, and in-unit data obtained from Soldiers, their supervisors, cadre, and peers, and archival sources at multiple points in time using a variety of data collection methods. The Cycle 17 full sample includes 1,132,118 applicants who took the TAPAS; however, some of them did not access into the Army or were ineligible for inclusion in the analyses based on their education status, AFQT score, component, or service history.

After excluding 09C/09L, Education Tier 3, AFQT Category V, and prior service applicants from the full sample, the remaining 1,048,245 Soldiers were included in the TOPS Applicant Sample. This sample represents Soldiers who possess qualities that are most representative of applicants to the Army. A majority of the Soldiers included in the sample are listed as Regular Army, Education Tier 1; and were predominantly male, White, and non-Hispanic.

Additional analysis samples were created based on this initial sample; however, they include fewer Soldiers. Of the full Applicant Sample, 621,590 (59.3%) had a record in at least one of the administrative criterion data sources; 69,525 (6.6%) had IMT data collected from the schoolhouse and 8,513 (0.8%) had in-unit criterion data. The Applicant Sample and Validation Samples were used in subsequent analyses presented in Chapters 4 through 7, as well as their associated appendices.

CHAPTER 3: THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM (TAPAS)⁷

Stephen Stark, O. Sasha Chernyshenko, Christopher Nye, and Fritz Drasgow
(Drasgow Consulting Group)

Description

The TAPAS is a personality measurement tool originally developed by Drasgow Consulting Group (DCG) under the Army's Small Business Innovation Research (SBIR) program. The system builds on the foundational work of the Assessment of Individual Motivation ([AIM]; White & Young, 1998) by incorporating features designed to promote resistance to faking and by measuring narrow personality constructs (i.e., facets) that are known to predict outcomes in work settings. The TAPAS uses methods from item response theory (IRT) to construct and score items. It can be administered in multiple formats: (a) as a fixed length, *non-adaptive test* where examinees respond to the same sequence of items or (b) as an *adaptive test* where each examinee responds to a unique sequence of items selected to maximize measurement accuracy for that specific examinee.

The TAPAS uses an IRT model for multidimensional pairwise preference (MDPP) items (Stark, Chernyshenko, & Drasgow, 2005) as the basis for constructing, administering, and scoring personality tests that are designed to reduce response distortion (i.e., faking) and yield normative scores even with tests of high dimensionality (Stark, Chernyshenko, & Drasgow 2012). TAPAS items consist of pairs of personality statements for which a respondent's task is to choose the one that is "more like me." The two statements constituting each item are matched in terms of social desirability and statement location (extremity), and often represent different personality facets. This approach makes it more difficult for examinees to determine which answers are better from the Army's perspective, and thus it is harder to "fake good" on all facets throughout the course of a test than it is with single-statement Likert-type personality items. Stark et al. (2014) reported small mean differences in scores of individuals who might be motivated to increase their scores (i.e., Army applicants who were told that their score might affect their enlistment eligibility) compared to individuals not so motivated (Air Force applicants who were asked to complete the TAPAS for research purposes only). In short, the TAPAS' features make it more difficult for respondents to distort their responses to obtain more desirable scores.

The use of an IRT model also greatly increases the flexibility of the assessment process. A variety of test versions can be constructed to measure personality facets that are relevant to specific work contexts, and the measures can be administered via paper-and-pencil or computerized formats. If test content specifications (i.e., test blueprints) are comparable across versions, the respective scores can be readily compared because the metric of the statement parameters has already been established by calibrating response data obtained from a base or reference group (e.g., Army recruits). The same principle applies to adaptive testing, wherein each examinee receives a different set of items chosen specifically to reduce the error in his or

⁷ This chapter incorporates only minor edits from the one included in Knapp & Kirkendall (Eds.) (2018), *Tier One Performance Screen Initial Operational Test and Evaluation: 2015-2016 Biennial Report*.

her facet scores at points throughout the exam. Adaptive item selection enhances test security because there is less overlap across examinees in terms of the items presented.

Another important feature of the TAPAS is that pools of statements representing over two dozen narrow personality facets are available. The initial TAPAS trait taxonomy was developed using the results of several large-scale factor-analytic studies with the goal of identifying a comprehensive set of non-redundant narrow traits. Since then, additional facets have been added and these narrow traits, if necessary or desired, can be combined to form either the Big Five (the most common organization scheme for narrow personality traits) or any other number of broader traits (e.g., Integrity or Positive Core Self-Evaluations). This is advantageous for applied purposes because TAPAS versions can be created to fit a wide range of applications (both pre- and post-enlistment) and are not limited to a particular service branch or criterion. Selection of specific TAPAS facets can be guided by consulting the results of a meta-analytic study performed by DCG that mapped TAPAS facets to several important organizational criteria for military and civilian jobs (e.g., task proficiency, training performance, attrition) (Chernyshenko & Stark, 2007), as well as subsequent validation research. Table 3.1 presents the names of the TAPAS facets together with a description of a typical high scoring individual.

Table 3.1. TAPAS Facets 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.
Adventure Seeking	High scoring individuals enjoy participating in extreme sports and outdoor activities.
Aesthetics	High scoring individuals appreciate various forms of art and music and participate in art-related activities more than most people.
Attention Seeking	High scoring individuals tend to engage in behaviors that attract social attention. They are loud, loquacious, entertaining, and even boastful.
Commitment to Serve	High scoring individuals identify with the military and have a strong desire to serve their country.
Consideration	High scoring individuals are affectionate, compassionate, sensitive, and caring.
Cooperation	High scoring individuals are pleasant, trusting, cordial, non-critical, and easy to get along with.
Courage	High scoring individuals stand up to challenges and are not afraid to face dangerous situations.
Curiosity	High scoring individuals are inquisitive and perceptive; they are interested in learning new information and attend courses and workshops whenever they can.
Dominance	High scoring individuals are domineering, “take charge” and are often referred to by their peers as “natural leaders.”

Table 3.1. (Continued)

Facet Name	Brief Description
Even Tempered	High scoring individuals tend to be calm and stable. They don't often exhibit anger, hostility, or aggression.
Ingenuity	High scoring individuals are inventive and can think "outside of the box."
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.
Responsibility	High scoring individuals are dependable, reliable, and make every effort to keep their promises.
Self Control	High scoring individuals tend to be cautious, levelheaded, able to delay gratification, and patient.
Selflessness	High scoring individuals are generous with their time and resources.
Situational Awareness	High scoring individuals pay attention to their surroundings and rarely get lost or surprised.
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.
Virtue	High scoring individuals strive to adhere to standards of honesty, morality, and "good Samaritan" behavior.

Scoring details and the criterion-related validation work that led to the inclusion of TAPAS in the TOPS IOT&E can be found in the *Expanded Enlistment Eligibility Metrics* report (Knapp & Heffner, 2010) and in earlier evaluation reports in this series (e.g., Knapp & Heffner, 2011).

Psychometric Properties of TAPAS Test Versions

As part of the TOPS IOT&E, nine versions of the TAPAS have been administered (see Table 3.2). The different versions have allowed ARI to explore the value of alternative facets and to retire the statement pools that were exposed in research settings. Currently, MEPCOM testing uses a statement pool developed solely for use by U.S. military services. All versions created in August 2011 or later use DoD-owned statement pools. In the present report, the validation analyses reported in Chapter 5 use all TAPAS versions except for the first one (13D-CAT) in which some of the dimensions have minor conceptual dissimilarities from the same dimensions in later versions.

Table 3.2. 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
15D-CAT v4	October 2009 to August 2011	15	Yes	120
15D-CAT v5	August 2011 to September 2013	15	Yes	120
15D-CAT v7	August 2011 to September 2013	15	Yes	120
15D-CAT v8	August 2011 to September 2013	15	Yes	120
13D-CAT v9	September 2013 – present	13	Yes	120
13D-CAT v10	September 2013 – present	13	Yes	120
13D-CAT v11	September 2013 – present	13	Yes	120

As a test security measure, form equivalence information is provided in a limited distribution addendum. Scores have been standardized within TAPAS versions to enable cross-version analyses.

Descriptive statistics, subgroup score comparisons, and intercorrelations of individual TAPAS scale scores, TAPAS composite scores, and AFQT scores are provided in Appendix B. Previous evaluation reports (e.g., Knapp & Kirkendall, 2018) also reported correlations of TAPAS scales with the ASVAB subtests and Aptitude Area composites. Because most of the observed correlations between TAPAS scales and ASVAB subtests were in the -.20 to +.20 range during the initial stages of TAPAS development, the two measures are judged to provide non-redundant information about applicants' dispositions, which is advantageous in selection and assignment contexts.

TAPAS Composites

An initial Education Tier 1 performance screen was developed from the TAPAS-95s scales for the purpose of testing in an applicant setting (Allen et al., 2010).⁸ This was accomplished by (a) identifying key criteria of most interest to the Army, (b) sorting these criteria into “can do” and “will do” categories (see below), and (c) selecting composite scales corresponding to the can do and will do criteria, taking into account both theoretical rationale and empirical results. Two unit-weighted TAPAS composites were initially developed: (a) Can-Do (for predicting technical training performance and completion) and (b) Will-Do (for predicting attrition and motivation-based performance). These composites were used operationally from March 2011 – September 2013.

⁸ TAPAS-95s was a paper-and-pencil, static version of the TAPAS used in the Army Class research.

A subsequent set of composites was developed by DCG and includes three regression-weighted scores: (a) Can-Do, (b) Will-Do, and (c) Adaptation (for predicting attrition). These scores became available for Army decision-making in September 2013. More information about how these newer composites were developed is provided in a limited distribution addendum. Those interested in obtaining a copy of this addendum should contact the editors for further information. The specific facet scales comprising each TAPAS composite is also close-held information given the operational nature of this measure.

The criterion-related validation analyses in Chapter 7 use the composite scores developed in 2013. Not all versions of the earlier TAPAS versions included the scales comprising the Can-Do and Adaptation composites. This is reflected in substantially smaller sample sizes for analyses involving the Can-Do composites. To maximize analysis sample sizes, scores on the Adaptation composite were computed using modified formulas for TAPAS versions 7 and 8. Scoring details and evidence of score equivalence across formulae is documented in a separately published ARI research note.

Summary

The purpose of this chapter was to describe the primary predictor measure being evaluated in the TOPS IOT&E. The TAPAS is unique among personality measures because it uses forced-choice pairwise items and IRT to promote resistance to faking. Promising initial validation research conducted as part of EEEM has been followed by additional research showing the validity of TAPAS in operational settings (Nye et al., 2012).

CHAPTER 4: DESCRIPTION AND PSYCHOMETRIC PROPERTIES OF OBJECTIVE CRITERION MEASURES

Christopher Graves and William Taylor (HumRRO)

This chapter describes objective criterion measures that were 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 assessed Soldier performance primarily via job knowledge tests (JKTs) and the administrative criterion of Advanced Individual Training (AIT) course grades. 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). Performance in AIT, as reflected by course grades and training restarts (retaking a course typically due to poor performance), provides an efficient measure of early-career performance.

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)

The JKTs administered as part of the TOPS IOT&E included a general Warrior Tasks and Battle Drills (WTBD) JKT and 12 MOS-specific JKTs. The WTBD JKT was administered to all Soldiers participating in the research, and the MOS-specific JKTs were administered to Soldiers in selected MOS to provide a more job-focused assessment. At the outset of the TOPS project, MOS-specific JKTs were developed for seven high-density MOS. Development of these JKTs drew from item pools used in prior ARI projects, including Select21 (Collins, Le, & Schantz, 2005) and Army Class (Moriarty, Campbell, Heffner, & Knapp, 2009). In 2014, JKTs for five additional MOS were developed for the purpose of examining the impact of recent changes in personnel policy. 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 and were last reported on in the TOPS 2014 annual report (Knapp & Wolters, 2017). The TOPS IOT&E 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)
- Human Resources Specialist (42A)
- Combat Medic Specialist (68W)
- Motor Transport Operator (88M)
- Wheeled Vehicle Mechanic (91B)

JKT Development

Development of JKTs 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 et al., 2009). This report summarizes the process in three steps:

- Develop test blueprints
- Develop test items
- Create and finalize test forms.

Develop Test Blueprints

Analysts worked with consultants and Army subject matter experts (SMEs) to develop JKT blueprints. Typically, consultants provided input to produce draft blueprints, which were then reviewed and finalized by Army SMEs. Most Army SME reviews occurred during in-person workshops at the respective non-commissioned officer (NCO) Academies or at ARI. Some Army SME reviews were conducted by email and telephone. Throughout the blueprint development process, SMEs received training on the purpose, design, and development of JKT blueprints.

Blueprints were to cover, as completely and efficiently as possible, the full job domain of each MOS and the WTBD tasks, as applicable. Generally, tests were intended to be over-length, in that they would include approximately 1.5 times the number of items needed to provide a valid test of Soldier knowledge. The additional items provided a cushion that allowed items with poor item statistics to be dropped while retaining a sufficient number of items to provide a reliable measure of Soldier knowledge. A typical blueprint specified approximately 45 items for a test, which in its final version would include approximately 30 items.

The job domain for each MOS was those critical tasks contained in Soldier Training Publications (STPs). Our general course of action for developing blueprints was to identify overall performance areas, which reflected the subject areas by which critical tasks are organized in the STPs. We used only the subject areas containing skill level (SL) 10 tasks, which are directed at pay grades E-1 to E-4. Most of these tasks are taught in IMT and trained during a Soldier's first year in unit. In addition to STPs, we also relied on IMT course programs of instruction and lesson materials. After identifying performance areas, we then delineated more specific performance requirements in each performance area. SMEs provided guidance on how to condense extensive content into a workable set of requirements that encompassed key tasks.

Blueprint development began with construction of a draft blueprint, which was then evaluated by SMEs for accuracy, currency, and coverage. The SMEs were then asked to weight performance areas to determine the percentage of items to be included in each area, corresponding to the frequency and importance of task performance in each area. The SMEs then weighted the

performance requirements within each area by ranking the requirements in order of importance. Together, the weighting exercises allowed us to calculate the number of items required for each performance area and requirement. The resulting blueprint for each MOS was used to develop test items. Over the course of test development, blueprint specifications were occasionally modified to reflect the feasibility of generating usable items in specific areas.

Develop Test Items

All item writers received training on the item writing process, as well as on the criteria and guidelines for viable items. The processes for developing and reviewing items involved the production of draft items followed by a series of iterative internal and external reviews. Figure 4.1 shows the development process used to develop the JKTs for the five MOS added in 2014.

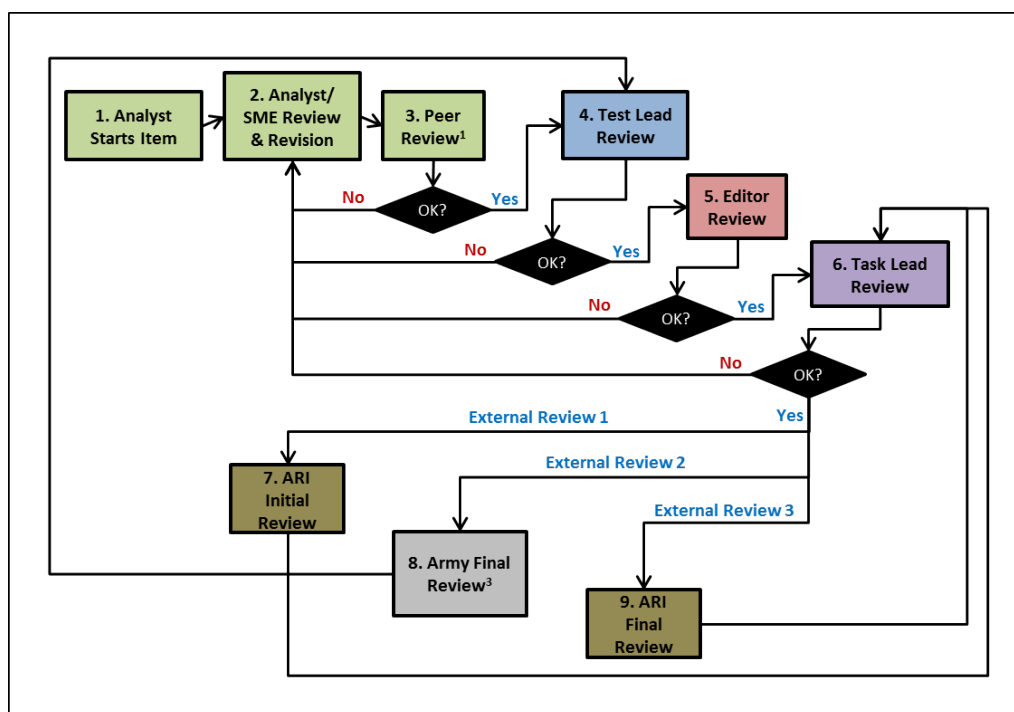


Figure 4.1. Job knowledge test item development and review process.

Most JKT items used a multiple-choice format with four response options. However, other formats, such as multiple-response (i.e., check all that apply), rank ordering, and matching were also used. Some items used images in addition to text to increase realism and reduce the reading burden on test-takers.

Create and Finalize Test Forms

Test forms are typically created in an initial form, piloted, and then amended into a final form based on pilot data. Piloting initial forms provides valuable information on internal consistency estimates, difficulty levels, and item discrimination (Moriarty et al., 2009). Use of these statistics enables the selection, elimination, and revision of items for scoring as required to craft more reliable and efficient measures.

Pilot testing for the TOPS JKTs was conducted in conjunction with the routine IOT&E data collections. The initial overlength test forms were administered until sample sizes were sufficient to yield reliable item statistics ($n > 100$). At this point, we reviewed item statistics to ensure the answer keys were programmed correctly and identify poorly performing items that required editing or removal. Considering the item statistics and test form reliability estimates, we finalized the test forms for future administration. In all cases, it was possible to calculate equivalent scores on both the pilot and final test forms.

Administration and Scoring

Test items could be worth multiple points, depending on item type. We computed a single, overall raw score for each JKT by summing the total number of points Soldiers earned across all items and computing a percent correct score based on the maximum number of points that could be obtained on each test. For subsequent analyses, we converted the total raw scores to standardized scores (or z-scores) by standardizing the scores *within* each MOS. A JKT score was flagged and not included in analysis if the Soldier (a) omitted more than 10% of the assessment items, (b) took fewer than five minutes to complete the entire assessment, or (c) selected an implausible response to one or more careless responding items which were included on most, but not all, of the JKTs (Knapp, Owens, & Allen, 2012).

Table 4.1 shows coefficient alpha reliability estimates for the WTBD and MOS-Specific JKTs. The MOS-specific reliability estimates were reasonably high (mean = 80.33). The WTBD reliability estimates were considerably lower (IMT = .68, in-unit = .56). It is possible that this could reflect greater heterogeneity in content for the WTBD versus the MOS-specific tests.

Table 4.1. Job Knowledge Test (JKT) Reliability Estimates in the IMT and In-Unit Validation Samples

Domain/JKT	<i>n</i>	<i>α</i>
IMT		
<i>MOS-Specific</i>		
Infantryman (11B/C/X + 18X)	12,459	.80
Combat Engineer (12B)	1,443	.79
Cannon Crewmember (13B)	159	.79
Field Artillery Tactical Data Systems Specialist (13D)	703	.76
Fire Support Specialist (13F)	830	.84
Cavalry Scout (19D)	1,492	.80
Armor Crewman (19K)	1,688	.81
Military Police (31B)	6,391	.81
Human Resources Specialist (42A)	1,183	.76
Combat Medic Specialist (68W)	7,721	.83
Motor Transport Operator (88M)	5,087	.77
Wheeled Vehicle Mechanic (91B)	1,684	.88
<i>WTBD (Army-Wide)</i>	56,919	.68
In-Unit		
<i>WTBD (Army-Wide)</i>	7,073	.56

Note. α = Coefficient Alpha. WTBD = Warrior Tasks and Battle Drills.

Table 4.2 shows descriptive statistics for the IMT MOS-specific JKTs and for the IMT and in-unit WTBD JKTs. The IMT JKT scores averaged roughly 62% correct. In-unit WTBD JKT scores were very similar, with a rough average of 61% correct. The correlation between the IMT MOS-specific JKTs and the WTBD JKT was .54, indicating that they were related but provided unique information about Soldier performance.

Table 4.2. Job Knowledge Test (JKT) Descriptive Statistics in the IMT Validation Sample

Domain/JKT	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>	<i>r</i> _{WTBD}	<i>r</i> _{AFQT}
IMT							
<i>MOS-Specific</i>							
Infantryman (11B/C/X + 18X)	12,051	59.14	11.00	23.33	86.67	0.62	0.43
Combat Engineer (12B)	1,394	57.69	16.21	8.57	97.14	0.58	0.52
Cannon Crewmember (13B)	154	57.49	15.80	13.89	91.67	0.55	0.30
Field Artillery Tactical Data Systems Specialist (13D)	684	50.31	14.94	8.11	81.08	0.63	0.41
Fire Support Specialist (13F)	807	63.18	17.44	14.71	97.06	0.60	0.32
Cavalry Scout (19D)	1,446	48.11	13.29	13.79	75.86	0.58	0.33
Armor Crewman (19K)	1,663	66.77	12.95	25.42	96.61	0.56	0.28
Military Police (31B)	6,218	64.20	9.71	25.44	91.26	0.56	0.42
Human Resources Specialist (42A)	1,171	55.07	12.56	15.09	86.79	0.56	0.43
Combat Medic Specialist (68W)	7,505	71.78	9.10	23.47	92.86	0.46	0.27
Motor Transport Operator (88M)	4,941	62.45	9.90	25.71	88.89	0.57	0.41
Wheeled Vehicle Mechanic (91B)	1,646	56.70	12.58	23.40	90.72	0.51	0.30
All MOS Combined	39,680	62.30	12.62	8.11	97.14	0.54	0.35
<i>WTBD (Army-Wide)</i>	55,994	61.58	12.50	6.45	97.30		0.41
In-Unit							
<i>WTBD (Army-Wide)</i>	7,063	62.17	10.67	15.38	96.15		0.44

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 the TOPS IOT&E included attrition, number of training restarts, and Advanced Individual Training (AIT) grades.

Attrition

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. Separation data are reported for Regular Army Soldiers only because reliable data for the Reserve and National Guard components were not available. Evaluation analyses cover attrition through 24 months of service. Table 4.3 summarizes basic descriptive statistics for attrition.

Training Restarts

Soldiers' IMT completion status and whether they graduated from IMT with training recycles or restarts were extracted from ATRRS (Army Training Requirements and Resources System). IMT recycles identify Soldiers who had to begin training again during IMT. IMT restarts identify Soldiers that graduated IMT with at least one failure (i.e., failed a component of training). IMT restarts were further divided into failures due to academic reasons versus those that were due to disciplinary, medical, or other reasons. Soldiers who had not had an opportunity to complete their IMT at the time data were extracted were excluded from analyses. Table 4.3 presents the base rates of Soldiers with at least one training recycle or restart during IMT.

Table 4.3. 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>
12-Month	298,531	43,519	14.6
24-Month	240,983	51,696	21.5
<i>Restarted Initial Military Training (IMT)</i>	<i>n^b</i>	<i>n_{Restarted}</i>	<i>%Restarted</i>
IMT Recycles	618,852	32,726	5.3
IMT Restarts	375,275	53,271	14.2
For Pejorative Reasons	369,503	47,204	12.8
For Academic Reasons	361,588	39,523	10.9

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 ($n < 15$).

AIT Grade

Final AIT Grade represents the cumulative grade across all courses administratively recorded for the Soldier in the RITMS (Resident Individual Training Management System) database. We computed a standardized version of Final AIT Grade by standardizing each course grade for courses with 15 or more Soldiers. We have not been able to access RITMS data since 2014, so this variable is only available for older cases in the TOPS database. Note that final grades from One Station Unit Training (OSUT) courses are excluded from the data file because the variance in grades is highly restricted or based on a pass-fail metric that is redundant with data from ATRRS. Table 4.4 presents basic descriptive statistics for AIT grades.

Table 4.4. Base Rates for Advanced Individual Training Grades in the Validation Sample

<i>Final AIT Grades</i>	<i>n^a</i>	<i>M</i>	<i>SD</i>
Overall Average (Unstandardized)	42,143	91.67	8.18
Overall Average (Standardized within MOS)	41,845	0.06	0.80

Note. AIT = Advanced Individual Training. MOS = Military Occupational Specialty. *n^a* = number of Soldiers with AIT grade data at time data were extracted.

Summary

In this chapter, we have described tests of job knowledge and administrative data that provided objective indices of important aspects of Soldier success in the TOPS IOT&E. The JKTs included a series of tests specific to selected MOS as well as an Army-wide test of knowledge of warrior tasks and battle drills. Creation of these tests involved working with SMEs to develop test blueprints and items, piloting test items, and creating and finalizing test forms. Results showed that most of the tests were sufficiently reliable, with an average alpha of 80.33 across all MOS-specific tests, and differentiating of performance, with an average IMT score of 62% and reasonable score variability on the MOS-specific tests. Administrative data included attrition, the number of training recycles and restarts, and final AIT grades (when such were available), which provided additional indices of Soldier performance.

CHAPTER 5: ARMY LIFE QUESTIONNAIRE

Colby K. Nesbitt (HumRRO), Elizabeth Salmon (ARI), and Cristina D. Kirkendall (ARI)

Background

The Army Life Questionnaire (ALQ) was used in the TOPS IOT&E to collect self-reported administrative data and attitudes. The questionnaire asked 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). There are two versions of the ALQ: one for administration to Soldiers during training (IMT) and one for Soldiers in their units of assignment.

The ALQ was originally developed under the Select21 project (Knapp, Sager, & Tremble, 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. Because the ALQ is used for multiple ARI research efforts, some additional content was added for purposes other than the TOPS IOT&E. The revisions not directly related to the TOPS IOT&E will only be briefly described.

Self-Report Administrative Criteria

Originally, self-report administrative criteria were collected to supplement administrative personnel records that were relatively inaccessible and of questionable accuracy. The ALQ was initially used to gather self-report administrative information in the following areas: (a) Awards, Decorations, and Achievements; (b) Disciplinary Actions (e.g., Article 15s); (c) performance in training (e.g., training restarts); and (d) test scores (e.g., Army Physical Fitness Test [APFT]).

The APFT, which Soldiers take 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. Disciplinary Incidents were scored two ways – first, 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) and second, as a continuous count of the number of associated items for which there was an affirmative response. On the IMT ALQ, Training Restarts was a dichotomous variable in response to an item asking whether the Soldiers ever had to restart training and Training Achievement was a composite variable in response to two questions about training achievements (reduced to a single question in the most recent revised ALQ).

While self-reported administrative outcomes can be problematic, previous research has found acceptable reporting rates. For example, during pilot testing of the *Personnel File Form* in Project A (predecessor of the self-report variables in the ALQ), Soldiers reported higher numbers of both positive (e.g., awards) and negative (e.g., Article 15s) indicators compared to official records (Riegelhaupt, Harris, & Sadacca, 1987).

Revisions to ALQ Attitudinal Scales

Throughout the TOPS IOT&E research, the ALQ contained items rated on Likert-type scales to tap content related to satisfaction and fit. The scales differed somewhat between the IMT and in-unit versions, but captured constructs thought to be predictive of retention behavior (e.g., Affective Commitment, MOS Fit, Army Life Adjustment). In 2016, we undertook an effort to revise these scales to accommodate new content (discussed in the next section) without appreciably increasing total administration time. Where possible, we sought to maximize consistency across the IMT and in-unit ALQs. Two scales, Attrition Cognition and Normative Commitment, were dropped entirely because they showed considerable empirical and conceptual overlap with other ALQ scales.

Item and Scale Reduction

The goal here was to identify a subset of items within each scale (e.g., Army Life Adjustment) that would capture the construct domain with a similar degree of reliability. We performed analyses of item content and statistics using the database from the 12th TOPS IOT&E cycle. ($n = 44,743$ - $66,357$ for IMT and $n = 6,320$ - $10,408$ for in-unit). For the content analysis, we identified item content that was redundant, awkwardly worded, or non-critical. Items were also considered for removal based on item-level statistics, including extremely high or low mean, extremely high endorsement of a single response option, low variance, low item-total correlation, and high scale coefficient alpha if item removed. We performed scale revisions iteratively, with consideration for both content representation and psychometric properties of the revised scales.

Pilot Study

The reduced scales were administered to in-unit Soldiers between fall of 2016 and winter of 2017 as part of a pilot study. Those data ($n = 741$ - 794) were used to evaluate the revised scale properties. Table 5.1 contains the background and demographic information for the sample.

Table 5.1. Background and Demographic Characteristics of Pilot Sample

	<i>n</i>	Percent
Gender		
Male	559	70.4
Female	121	15.2
Missing	114	14.4
Ethnicity		
White, Not Hispanic	290	36.5
Black, Not Hispanic	183	23
Hispanic	135	17
Asian	38	4.8
Other	57	7.2
Missing	91	11.5

We inspected the item-level statistics to ensure items were performing as expected. Scale reliability estimates were computed and are presented in Table 5.2. We observed that the coefficients alpha of the shortened scales decreased only slightly and remained in the acceptable range of satisfactory internal consistency.

Table 5.2. Reliability Estimates of ALQ Scales, Before and After Revision

Existing Scales	Before Revision (TOPS Database)		After Revision (TOPS Database)		Pilot Study	
	Alpha	# Items	Alpha	# Items	Alpha	# Items
Affective Commitment	0.83	7	0.78	3	0.85	3
Army Fit	0.89	6	0.78	4	0.79	4
Army Career Intentions	0.93	3	-	1	-	1
Army Reenlistment Intentions	0.80	4	-	1	-	1
MOS Fit	0.93	9	0.86	4	0.85	4
MOS Satisfaction	0.93	6	0.87	3	0.86	3

Note. Based on the TOPS Cycle 16 database. IMT ($n = 53,100-53,945$) and in-unit ($n = 7,276-7,809$) data combined. Excludes single-item scales.

TOPS Database Analyses

We performed a final comparison of the old and new versions of existing ALQ scales using the first database in which these changes were incorporated, the 16th TOPS IOT&E cycle ($n = 53,100-53,945$ for IMT and $n = 532-7,809$ for in-unit). This analysis included computing the means and standard deviations of the old and new versions of the scale, as well as estimating the correlations between the old and new scores. As shown in Table 5.3, the correlations between the new and old versions ranged from 0.89 and 0.97. Based on the accumulated evidence from this and the prior analyses, the research team concluded that the reduced scales showed a high degree of consistency with the original set of measures.

Table 5.3. Descriptive Statistics and Correlations between Scores on Existing ALQ Scales, Before and After Revision

	New Version		Old Version		<i>r</i>
	Mean	<i>SD</i>	Mean	<i>SD</i>	
IMT Scales					
Affective Commitment	3.75	0.81	3.86	0.70	0.94
Army Career Intentions	3.55	1.14	3.26	1.10	0.93
Army Fit	4.00	0.68	4.05	0.62	0.94
Army Life Adjustment	4.02	0.81	4.02	0.71	0.94
Army Reenlistment Intentions	3.65	1.09	3.52	0.97	0.89
MOS Fit	3.84	0.81	3.73	0.82	0.96
In-Unit Scales					
Affective Commitment	3.26	0.93	3.49	0.82	0.94
Army Career Intentions	2.80	1.33	2.55	1.21	0.95
Army Fit	3.58	0.94	3.81	0.74	0.93
Army Reenlistment Intentions	3.12	1.40	2.94	1.18	—
MOS Fit	3.37	0.95	3.20	0.93	0.96
MOS Satisfaction	3.40	0.96	3.43	0.92	0.97

Note. Based on the TOPS Cycle 16 database. IMT ($n = 53,100-53,945$) and in-unit ($n = 7,276-7,809$).

Development of New ALQ Content

Identification of New Criterion Constructs

New criterion constructs were selected for inclusion on the ALQ based on a review of previous Army criterion measures, their relevance to examining the potential of the TAPAS and other predictor measures to improve new Soldier selection and MOS assignment, and Army priorities. Addition of these 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. The following five constructs were chosen for this purpose: (a) Organizational Citizenship Behavior (OCB), (b) Leadership Behavior, (c) Motivation to Lead (MTL), (d) Counterproductive Soldier Behavior (CSB), and (e) Resilience. Because the ALQ is used for ARI projects other than the TOPS IOT&E, four constructs were added that are not directly related to this project, including Work-Life Balance, Deployment Satisfaction, Army Choice, and MOS Choice. Table 5.4 summarizes the construct definitions of the chosen domains.

Content Generation

For each of the new constructs, we reviewed the academic literature and prior military research and compiled a comprehensive database of relevant item content from existing validated measures, where applicable. Notably, most of the available instruments included in academic research were intended for civilian samples. Accordingly, content was adapted to reflect the context and level of the Army Soldier population. Items that were irrelevant or redundant were excluded.

Table 5.4. New Target Content for the Revised ALQ

Construct	Definition
Counterproductive Soldier Behavior (CSB)	Intentional behaviors that harm or are intended to harm another Soldier or the legitimate interests of the unit.
Motivation to Lead (MTL)	The factors that affect individual's decisions to assume leadership training, roles, and responsibilities, and affect his or her intensity of effort at leading and persistence as a leader (Chan & Drasgow, 2001).
Organizational Citizenship Behavior (OCB)	Engaging in voluntary behaviors to help another individual or the organization itself (Bateman & Organ, 1983).
Leadership Behavior	Behaviors that Soldiers engage in to display leadership qualities, absent of an official leadership role.
Resilience	The capacity to overcome difficult life events with minimal disruption or long-term negative impacts on psychological and physical functioning (Bonanno, 2004).
Work-Life Balance	The extent to which Soldiers are satisfied with the balance between their Army life and family life.
Deployment Satisfaction	The self-reported career and affective outcomes resulting from deployment.
Army Choice	Reasons the Soldier chose to join the Army and which individuals influenced that decision.
MOS Choice	Reasons the Soldier chose their initial entry MOS, which individuals influenced that decision, and the Soldier's desire to reclassify into another MOS.

New ALQ Scales Included in TOPS IOT&E

Counterproductive Soldier Behavior

The research team was interested in collecting information on counterproductive behaviors beyond the questions about disciplinary incidents already included in the ALQ. We consulted relevant studies and existing measures of counterproductive work behavior (Gruys & Sackett, 2003; Robinson & Bennett, 1995; Spector, Fox, Penney, Bruursema, Goh, & Kessler, 2006). Much of the content from measures used in academic research was not germane to the Army context due to the severity of the behaviors or characteristics of the work environment. After a review by the research team to remove redundant or irrelevant behaviors, an initial set of counterproductive behaviors was reviewed by an Army subject matter expert (SME). The behavior list was further edited to retain the most relevant content and adapt the behaviors to the Army context. The list of 22 items represented a wide range of behaviors directed at both fellow Soldiers and the Army as a whole. These behaviors represented the following dimensions: (a) withdrawal and poor attendance, (b) waste, (c) quality and safety, (d) destruction, (e) covering up errors or mistakes, (f) insubordination, (g) interpersonal abuse, (h) theft, and (i) social media use.

Motivation to Lead

Content for this domain was adapted from a measure developed by Chan and Drasgow (2001) and administered in a military setting by Amit, Lisak, Popper, and Gal (2007). Their conceptual and empirical model of MTL includes three underlying dimensions: Affective, Noncalculative, and Social-Normative. The Affective subscale measures an individual's inherent preference to be the leader of a group. The Noncalculative subscale assesses a person's desire to be leader of a group due to the benefits or privileges afforded to leaders. Finally, the Social-Normative subscale captures the desire to lead because of a sense of duty. Items from across those three dimensions were selected and/or adapted based on their relevance to the Army context. The list included six, five, and six items for each subscale, respectively.

Organizational Citizenship Behavior

Our goal was to measure contextual behaviors beyond task performance. For content in this area, theory and measures from several studies of contextual and organizational citizenship behavior were examined (Dalal, Lam, Weiss, Welch, & Hulin, 2009; Motowidlo & Van Scotter, 1994; Podsakoff, Ahearne, & MacKenzie, 1997). Content areas included behaviors such as teamwork, rule adherence, and volunteerism, among others. Ultimately, the team developed a list of 13 items that represented a variety of behaviors (e.g., supporting fellow Soldiers, volunteering for additional duties) directed toward the Soldier's job roles, other Soldiers, the Soldier's unit, and the Army as a whole.

Leadership Behavior

We consulted popular leadership theory to develop content for the Leadership Behavior scale. In particular, we relied on the two-factor framework of consideration and initiating structure (Yukl, Gordon, & Taber, 2002), and identified relevant subdimensions such as demonstrating support/concern, conflict management, and monitoring operations, among others. We generated a list of 14 behaviors. It is important to note that the Soldier population taking the ALQ were

generally not at a supervisor level; accordingly, content germane to supervisory leadership rather than peer leadership (e.g., evaluating subordinates) was excluded or adapted.

Resilience

The research team considered aspects of resilience that support effective functioning in stressful situations, with particular focus on the components of resilience that have been linked to success in military contexts. The measure was designed to assess as many core resilience topics as possible, including various active and passive strategies, coping efficacy, and social support, but also minimized the number of items included due to time constraints. Measure development began with a survey of the literature and existing measures of resilience. An initial examination of these measures removed duplicate items and constructs that were clearly not relevant in the Army context. With the remaining scales, we catalogued a set of approximately 53 items and noted the dimension or sub-dimension of resilience that each was designed to measure. The resulting list was sorted by dimension to assess the coverage of each construct space. The number of items in each dimension was then reduced to a total of 16 to ensure a reasonable administration time. Deficiencies were identified in a few areas, including physical coping strategies, so we drafted new items to provide construct coverage.

New ALQ Scales Not Included in TOPS IOT&E

Work/Life Balance

This scale was added to the ALQ due to its potential impact on retention and attrition. Work/Life Balance may also be related to resilience. For this scale, we utilized previous ARI criterion instruments reported in Strickland (2005) to adopt or adapt items for this domain. The final scale consisted of five items.

Deployment Satisfaction

As with Work/Life Balance, researchers utilized previous ARI criterion instruments to adopt items for this domain. Deployment satisfaction was included on previous versions of the ALQ (e.g., Moriarty, Campbell, Heffner, & Knapp, 2009) but had been removed when Army deployment frequency decreased. Items used on the previous ALQ were added back in for the new version. The scale consisted of between one and eight items, depending on whether the Soldier had been deployed with the Army.

Army Choice

Two questions were added to the ALQ to assess what factors and individuals influenced the Soldier's decision to join the Army. Response options were adopted/adapted from previous Army measures. Soldiers had the choice of several reasons for joining and potential influencers and were asked to choose all that apply. Soldiers were also given the choice to indicate "Other" and write in a response.

MOS Choice

Several questions were added to address whether the Soldier came into the Army with a particular MOS in mind. This scale also asked what factors and individuals influenced the Soldier's MOS choice. Soldiers were also asked if they would reclassify to another MOS if given the opportunity and why they would choose to leave their current MOS. The number of items presented to each Soldier varied from 4 to 13 depending on previous item responses.

Response Scale

Several of the new constructs assessed by the ALQ, most notably OCB and CSB, are prone to social desirability effects. The literature in these areas lays out best practices for measurement, such as using response scales based on frequency rather than agreement (Dalal, 2005). The team took these recommendations into account while also balancing the reality that many Soldiers may not have the opportunity to engage in OCB and/or CSB, particularly during IMT. For this reason, the team developed a prompt for the OCB and CSB scales ("Compared to other Soldiers in your class [or in your unit], how often do you...") and a comparative frequency response scale (e.g., 1 = Far less often; 7 = Far more often).

Subject Matter Expert Review

The initial review of existing, validated measures yielded a satisfactory pool of items for measuring MTL. The content of the existing measures was sufficiently generalized and non-context-specific, making it relatively easy to adapt to the Army context. Developing or adapting existing content for the more behaviorally-oriented scales (i.e., Leadership Behavior, OCB, and CSB), however, required further refinement from Army SMEs. It was particularly critical to gather information on the opportunity to engage in these behaviors and their relevance to Soldiers at the level of interest. Approximately 20 NCOs from various MOS discussed the nature of those performance areas and reviewed the draft items. The SMEs provided input on the relevance and opportunity to demonstrate the specified behaviors and identified a few additional relevant behaviors to be included in the scales (e.g., social media behaviors to consider for CSB). Conversations with SMEs also yielded insights about the likelihood of response bias. The content for the resilience measure was reviewed by a different group of Army SMEs, who edited items and rated them on their importance in the training context.

Pilot Study

The new ALQ content was administered to a sample of Soldiers in the same pilot study and sample ($n = 741-794$) previously described. Deployment Satisfaction, Work/Life Balance, Army and MOS Choice were not included in the pilot as this content had been previously administered in the Army context. These data were used to conduct item- and scale-level analyses of the new measures. Specifically, we identified items that had extremely high or low means, low variability, extreme overlap with other items in the scale, low item-total correlations or high coefficients alpha with item removed. We also conducted confirmatory factor analyses and flagged items with low factor loadings or high cross-loadings. We discuss the results of the item and scale analyses for each scale below.

Counterproductive Soldier Behavior

Item analyses revealed that many items yielded means and variabilities that were too low to be useful, particularly for the social media behaviors. Accordingly, the social media subset of behaviors was excluded from the final version of the instrument. The structure of counterproductive work behaviors is the subject of debate in the academic literature. Accordingly, we investigated the factor structure to determine whether factor scores were sufficiently related to collapse into a single score. Based on the full set of items, bivariate correlations among the nine subscales ranged from 0.35 to 0.72, with an average correlation of 0.48. Based on this information and factor analyses, we determined that a single factor model was sufficient to represent the domain. Further, from a reporting standpoint, a single scale was more useful than a number of separate scales. Items with the highest base rates and variabilities were retained, with an effort to represent a diversity of behaviors.

Motivation to Lead

Factor analyses and subscale correlations suggested that the data conformed to the three-factor conceptual model of MTL. Moderate correlations and conceptual differences between the three subscales indicated that the subscores should be reported separately. The scales were revised to eliminate poorly performing items and to reduce redundancy.

Organizational Citizenship Behavior and Leadership Behavior

Before any items were dropped from the scales, the correlation between the Leadership Behavior and OCB scales was 0.82. This correlation, nearly as high as the coefficients alpha of these scales (0.92 for each scale), suggested that these two scales might be empirically indistinguishable. Although steps had been taken to distinguish the two constructs during item development, the content of these scales also overlapped. The Consideration and Initiating Structure components of leadership, when translated into a non-supervisory capacity, looked very similar to the Citizenship behaviors directed at fellow Soldiers and the Army itself. Accordingly, items from the two scales were combined and analyzed as a single unidimensional scale called OCB/Leadership (OCB/L).

Resilience

The Resilience scale item analyses revealed that several items yielded high means and low variability. Further review of the item correlations suggested that several items were redundant, and results of an exploratory factor analysis showed that most items loaded highly on a single factor. Based on these analyses, we pruned the items to remove problematic and redundant items and to preserve unidimensionality.

ALQ Composite Scores

Original Composite Scores

Previous TOPS evaluations used two ALQ composites that needed to be adjusted with the revised ALQ: Commitment and Fit and Retention Cognitions. Some items and associated scales that had previously comprised these composites were removed from both the IMT and in unit

ALQ. Although these changes made revisions necessary for both composites, our goal was to keep each as similar as possible to their original versions. We performed the work to revise the composites as part of the 16th TOPS IOT&E analysis cycle.

With respect to the Commitment and Fit composite, changes were made to each of the associated scales. For both IMT and in-unit, items were removed from Affective Commitment, Army Fit, and MOS Fit; and new items were added to MOS Fit. Most notably, the Normative Commitment scale was removed entirely from the revised ALQ. Thus, the revised IMT and in-unit Commitment and Fit composites were computed as the average of the revised Affective Commitment, Army Fit, and MOS Fit scales.

Changes to the Retention Cognitions composite included items being removed from the Career Intentions and Reenlistment Intentions scales at both IMT and in-unit. The items removed from the in-unit Reenlistment Intentions scale were replaced by a single item, which had been administered on the IMT measure but had not appeared on the previous in-unit questionnaire. Finally, the Attrition Cognitions scale was removed entirely from the ALQ. Given these changes, the revised IMT and in-unit Retention Cognitions composites reflect the average of the revised Career Intentions and Reenlistment Intentions scales. Table 5.5 compares the composition of both the original and revised Commitment and Fit and Retention Cognitions composites.

Table 5.5. ALQ Composites and Associated Scales for the Original and Revised Versions (IMT and In-Unit)

Scale	Commitment and Fit		Retention Cognitions	
	Original	Revised	Original	Revised
Affective Commitment	X	X		
Army Career Intentions			X	X
Army Fit	X	X		
Army Reenlistment Intentions			X	X
Attrition Cognition			X	
MOS Fit	X	X		
Normative Commitment	X			

Note. Army Reenlistment Intentions is included in both the original and revised composite, but the item that makes up the revised scale was not part of the original scale.

To evaluate comparability of the original and revised composites, we computed correlations between each version and the TAPAS scales as well as the supervisor performance ratings overall composite (described in Chapter 6). Then, we computed the difference between the correlations associated with the original and revised composites. This comparison allowed us to examine the extent to which the original and revised ALQ composite versions exhibited different relationships with other variables. Results of this analysis are shown in Table 5.6. Looking across Commitment and Fit and Retention Cognitions, correlational differences for the original and revised versions were small. For IMT, the average difference was $\Delta r = .01$ for the TAPAS scales. For in-unit, the average difference was $\Delta r = .00$ for the TAPAS scales.

Table 5.6. Differences in Correlations for the Original and Revised ALQ Composites

IMT						
Variable	Commitment and Fit			Retention Cognitions		
	Original	Revised	Δr	Original	Revised	Δr
<i>TAPAS Scales</i>						
Achievement	.11	.12	.00	.06	.07	.01
Adjustment	.02	.02	.00	-.01	-.01	.01
Attention Seeking	.03	.04	.01	-.03	-.01	.02
Commitment to Serve	.15	.19	.03	.22	.21	-.01
Cooperation	.02	.02	.00	.00	.00	.00
Courage	.10	.11	.00	.03	.05	.02
Dominance	.08	.09	.01	.04	.05	.01
Even Tempered	.02	.02	.00	.02	.02	.00
Intellectual Efficiency	.04	.04	.01	.02	.03	.01
Non-Delinquency	.04	.05	.00	.01	.01	.00
Optimism	.06	.07	.01	.01	.02	.00
Order	.00	.01	.01	.04	.03	-.01
Physical Conditioning	.04	.05	.00	-.02	-.01	.01
Responsibility	.09	.08	.00	.01	.02	.00
Self-Control	.02	.02	.00	.03	.02	.00
Selflessness	.05	.05	.00	.04	.05	.00
Situational Awareness	.06	.07	.01	.03	.04	.01
Sociability	.05	.07	.02	.05	.05	.00
Team Orientation	.03	.06	.02	.03	.04	.01
Tolerance	.03	.04	.00	.06	.06	.00
<i>Criterion Composites</i>						
IMT PRS-S Overall Perf Composite	.06	.05	.00	.01	.02	.02
In-Unit PRS-S Overall Composite	-.01	.00	.01	-.03	-.04	-.01
In-Unit						
Variable	Commitment and Fit			Retention Cognitions		
	Original	Revised	Δr	Original	Revised	Δr
<i>TAPAS Scales</i>						
Achievement	.07	.07	.00	.04	-.01	-.05
Adjustment	-.01	-.01	.00	—	—	—
Attention Seeking	.02	.02	.00	-.03	.11	.14
Commitment to Serve	.13	.12	-.01	.12	.19	.07
Cooperation	.03	.04	.00	.01	.18	.17
Courage	.06	.06	.00	-.07	.06	.13
Dominance	.03	.03	.00	.02	-.03	-.05
Even Tempered	.05	.05	.00	.02	.08	.06
Intellectual Efficiency	.00	.00	.00	-.03	.02	.05

Table 5.6. (Continued)

Variable	In-Unit					
	Commitment and Fit			Retention Cognitions		
	Original	Revised	Δr	Original	Revised	Δr
Non-Delinquency	.05	.05	.00	.04	.12	.09
Optimism	.07	.07	.00	.00	.13	.12
Order	-.01	.00	.00	.05	-.03	-.07
Physical Conditioning	.05	.04	-.01	-.01	.02	.03
Responsibility	.05	.04	-.01	.06	-.01	-.07
Self-Control	.03	.02	-.01	—	—	—
Selflessness	.05	.04	.00	.05	.06	.01
Situational Awareness	.07	.06	-.01	.07	.03	-.03
Sociability	.05	.04	.00	.03	-.04	-.07
Team Orientation	.03	.02	-.02	-.06	-.05	.01
Tolerance	.03	.04	.01	.07	-.02	-.09
<i>Criterion Composites</i>						
IMT PRS-S Overall Perf Composite	.09	.08	-.02	—	—	—
In-Unit PRS-S Overall Perf Composite	.14	.13	-.01	.02	.20	.18

Note. Based on the TOPS Cycle 16 database. For IMT correlations, $n = 415\text{--}39,495$. For IU correlations, $n = 123\text{--}4,504$. Correlations are not reported where $n < 100$.

New Composite Scores

With respect to the ALQ, six new scales were of interest for the TOPS IOT&E project: Motivation to Lead (MTL)-Affective, MTL-Noncalculative, MTL-Socionormative, Counterproductive Soldier Behavior (CSB), OCB/L, and Resilience.⁹ We began by reviewing the items and dimensions associated with each new scale. We used this qualitative review to propose meaningful *a priori* combinations of the scales and concluded that four of the scales represented relevant qualities for effective peer leadership. Specifically, we proposed a new Peer Leadership composite to include MTL-Affective, MTL-Noncalculative, MTL-Socionormative, and OCB/L. Correlations among the new scales also demonstrated expected theoretical relationships.

Confirmatory Factor Analysis Check

As a final step, we conducted a CFA to evaluate both the revised and new ALQ composites. Specifically, we fit a model that included each of the revised scales as well as the four new scales given that all were administered together on the ALQ. We fit separate models for the IMT and in-unit composites. In each model, the Commitment and Fit, Retention Cognitions, and Peer Leadership composites were included as correlated latent factors. Factor loadings for the scales provided support for the model, and both models fit the data well. Table 5.7 presents the fit statistics for these models. Given our results, we averaged the four new scales to create a new Peer Leadership criterion composite.

⁹ We did not include the new Work/Life Balance, Deployment Satisfaction, Army Choice, or MOS Choice scales in our analyses because they were not intended for use as criterion scores.

Table 5.7. CFA Model Fit Statistics for the Army Life Questionnaire (ALQ) New and Revised Criterion Composites

Model	χ^2	df	CFI	RMSEA	SRMR
IMT	2293.70	36	.99	.04	.04
In-Unit	177.67	24	.99	.03	.03

Final ALQ Contents and Associated Psychometric Properties

Table 5.8 shows the internal consistency reliability estimates for the IMT and in-unit ALQ scales and composites, all of which are at acceptable levels. Table 5.9 provides descriptive statistics for selected self-report administrative criteria, as well as the Likert-based scales and composites. Intercorrelations among the revised ALQ scales are shown in Appendix C (Table C.1). Appendix D shows items comprising the new and revised ALQ scales. We performed these analyses on the most recent database available (Cycle 17).

Table 5.8. ALQ Scale Reliability Estimates

Variable	IMT		In-Unit	
	Alpha	Number Scales/Items	Alpha	Number Scales/Items
Composites				
Commitment and Fit ^a	.81	3	.82	3
Peer Leadership ^a	.82	4	.84	4
Retention Cognitions	.92	2	.94	2
Likert Scales				
Affective Commitment	.80	3	.80	3
Army Fit	.77	4	.79	4
Army Life Adjustment	.79	4	—	—
Counterproductive Soldier Behavior	.85	9	.79	9
Deployment Satisfaction	—	—	.81	4
MOS Fit	.86	5	.88	5
MOS Satisfaction	—	—	.84	5
Motivation to Lead - Affective	.84	3	.84	3
Motivation to Lead - Noncalculative	.78	3	.78	3
Motivation to Lead - Social Normative	.77	3	.78	3
Organizational Citizenship Behavior/Leadership	.87	8	.87	8
Resilience	.81	7	.79	7
Work/Life Balance	.87	5	.86	5

Note. Army Career Intentions and Reenlistment Intentions are single-item measures, and therefore alpha is not reported for these variables.

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

Table 5.9 Revised ALQ Descriptive Statistics

Variable	IMT					In-Unit				
	<i>n</i>	Mean	<i>SD</i>	Min	Max	<i>n</i>	Mean	<i>SD</i>	Min	Max
Composites										
Commitment and Fit	63,144	3.83	0.65	1.00	5.00	8,153	3.40	0.77	1.00	5.00
Peer Leadership	17,012	3.29	0.45	1.00	5.00	895	3.25	0.49	1.00	5.00
Retention Cognitions	62,876	3.60	1.08	1.00	5.00	894	2.92	1.34	1.00	5.00
Likert Scales										
Affective Commitment	63,144	3.71	0.82	1.00	5.00	8,153	3.25	0.94	1.00	5.00
Army Career Intentions	62,943	3.55	1.14	1.00	5.00	8,117	2.79	1.33	1.00	5.00
Army Fit	63,144	3.96	0.69	1.00	5.00	8,153	3.57	0.94	1.00	5.00
Army Life Adjustment	63,144	3.94	0.84	1.00	5.00	—	—	—	—	—
Counterproductive Soldier Behavior	17,012	1.95	0.66	1.00	5.00	895	2.00	0.61	1.00	5.00
Deployment Satisfaction	—	—	—	—	—	243	3.36	1.03	1.00	5.00
MOS Fit	63,144	3.82	0.81	1.00	5.00	8,153	3.36	0.96	1.00	5.00
MOS Satisfaction	—	—	—	—	—	8,153	3.38	0.97	1.00	5.00
Motivation to Lead - Affective	17,012	3.46	0.82	1.00	5.00	895	3.42	0.87	1.00	5.00
Motivation to Lead - Noncalculative	17,012	2.27	0.86	1.00	5.00	895	2.37	0.91	1.00	5.00
Motivation to Lead - Social Normative	17,011	3.99	0.70	1.00	5.00	895	3.92	0.76	1.00	5.00
Organizational Citizenship Behavior/Leadership	17,012	3.45	0.68	1.00	5.00	895	3.29	0.74	1.00	5.00
Reenlistment Intentions	62,987	3.65	1.10	1.00	5.00	894	3.04	1.43	1.00	5.00
Resilience	17,010	3.97	0.62	1.00	5.00	895	3.87	0.66	1.00	5.00
Work/Life Balance	16,993	3.14	0.86	1.00	5.00	895	2.97	0.95	1.00	5.00
Administrative Criteria										
Disciplinary Incidents (#)	61,336	0.30	0.63	0.00	7.00	8,153	0.42	0.89	0.00	7.00
Disciplinary Incidents (Y/N)	61,336	0.23	0.42	0.00	1.00	8,153	0.26	0.44	0.00	1.00
APFT Score	60,798	252.12	28.21	180.00	300.00	7,796	253.01	28.22	180.00	300.00

Summary

The present effort sought to expand the criterion domain of behaviors and attitudes captured by the ALQ to increase the instrument's relevance and utility for future validations. Toward that end, the research team identified new content areas and developed psychometrically sound measures of those constructs. As we increased the breadth of content, we revised the original content to limit overall administration time. We identified conceptual and empirical redundancy within the existing scales. The result was revised measures that expanded the construct space and preserved the strong psychometric properties of the original instrument.

CHAPTER 6: DESCRIPTION AND PSYCHOMETRIC PROPERTIES OF PERFORMANCE RATING CRITERION MEASURES

Deirdre J. Knapp (HumRRO), Elizabeth Salmon (ARI), Cristina D. Kirkendall (ARI), and
Timothy Burgoyne (HumRRO)

Introduction

Performance ratings are desirable primarily because they can comprehensively cover technical, as well as motivational and other “non-cognitive” elements of performance. The latter is particularly important in the TOPS IOT&E, 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 policy-makers. 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.

The TOPS IOT&E research incorporated both supervisor (or IMT cadre) and peer ratings. Supervisor/cadre ratings are logistically somewhat easier to collect than peer ratings because it is easier to identify suitable raters. While obtaining ratings from multiple supervisor/cadre would improve rating score quality, it is difficult to obtain ratings from multiple supervisor/cadre raters who are sufficiently familiar with Soldier performance. Moreover, the time required for senior personnel to complete the ratings can be prohibitive. For IMT cadre, who might be asked to rate 30 or more trainees, the issue is both time and the ability to differentiate detailed performance aspects among so many Soldiers. These circumstances led to the introduction of peer ratings fairly late in the TOPS IOT&E program.

Prior to the introduction of peer ratings, there were multiple efforts to improve the psychometric quality of the supervisor rating scales which will be described further in this chapter. These efforts met with only minor improvements in measurement quality. To put these efforts in context, and support thinking about future criterion development efforts, this chapter ends with a brief historical overview of Army validation research incorporating performance ratings dating back to Project A (Campbell & Knapp, 2001).

The remainder of this chapter first describes the TOPS Performance Rating Scales for supervisors and cadre (PRS-S), starting with the IMT versions followed by the in-unit scales. Next, the recently introduced peer ratings (PRS-P, both IMT and in-unit) are described. The historical summary section referenced above will then help lead into a final summary and conclusions section.

Supervisor Performance Rating Scales

Background

The PRS-S, like the JKTs, were adapted from or based on previous research dating back to Project A (Moriarty, Campbell, Heffner, & Knapp, 2009). We developed Army-wide rating scales (IMT and in-unit) suitable for rating all Soldiers. We also developed rating scales covering MOS-specific content for selected MOS. In an effort to improve the quality of the ratings, we

revised the IMT scales early in the IOT&E (2011). At that time, we suspended collection of in-unit MOS-specific rating scales altogether to allow supervisors to focus on making higher quality ratings on the Army-wide scales.

All versions of the PRS-S included 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. Initially we used a 3-point familiarity scale, but this was changed to the following 4-point scale in 2011 to enable raters to more precisely depict their ability to judge each Soldier’s performance.

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.

IMT cadre completed their Soldier ratings without any research staff present whereas in-unit supervisors completed the ratings in a proctored environment. The online instrument included an instructional screen and in-unit raters could query a proctor if they had a question. The instructional screen listed the behavioral dimensions to be rated and encouraged raters to rate typical performance, carefully reading the name and definition of each dimension, and distinguishing performance across scales. Moreover, the following reminders appeared on each rating screen:

- Read the whole description
- Base your rating on what you have seen. Use Not Observed/Cannot Rate when necessary
- Reflect on Soldiers’ typical performance including strengths and weaknesses

Raters evaluating multiple Soldiers rated each Soldier on a given performance dimension before moving to the next dimension.

IMT PRS-S

Development

The original TOPS Army-wide PRS used the same performance dimensions that had been used in the Army Class research program (Moriarty et al., 2009). Those eight dimensions were identified and defined by (a) reviewing Army trainee rating dimensions identified in three prior research efforts, (b) considering available dimension-level interrater reliability evidence, and (c) Project A factor analysis results.

The Army Class instrument used a rating format in which multiple behavioral statements were written for each performance dimension, each describing either very high or very low levels of performance. These bipolar statement pairs were used to anchor 5-point rating scales, up to nine of which comprised the ratings for each performance dimension. A small group of senior NCO SMEs from a mix of MOS reviewed wording of all content in the Army Class rating scales and revisions were made as needed.

For TOPS, we converted the behavioral statements to high, medium, and low descriptive anchors for a more traditional 7-point Behaviorally Anchored Rating Scales (BARS) format, but the rating dimensions remained the same.

Depending on the MOS, the TOPS IMT MOS-specific rating scales were either updated from prior research (e.g., Army Class) or developed from scratch. Rating scale development and review was carried out in conjunction with development of the MOS-specific JKTs (described in Chapter 4) and involved multiple visits with IMT cadre to develop and review content. We followed processes employed over the course of similar ARI-sponsored criterion development efforts (e.g., Army Class). This included identifying and defining rating dimensions and conducting iterative Army SME reviews of content (dimension definitions and scale anchors).

To develop draft scales, analysts worked closely with paid consultant SMEs to identify and then define initial sets of rating dimensions for each of the PRS. After internal review, we presented the scales to Army SMEs (IMT instructors and platoon sergeants) during each of three workshops for each MOS. In the first workshop, SMEs (a) reviewed and refined the IMT scale dimensions and their definitions, (b) refined the in-unit rating dimensions, and (c) identified behavioral examples for each of three levels of performance: below expectations, meets expectations, and exceeds expectations. We asked SMEs the following questions:

- Have we included all relevant dimensions?
- Are the dimensions appropriate for the specified environment (i.e., IMT versus in-unit)?
- Have we defined the dimensions properly?
- Are the behavior examples appropriate?
- Are there behavior examples that we should add?

In subsequent workshops, we obtained review and revision input from additional Army SMEs. We finalized the rating scales based on these iterative reviews.

The scales used a traditional 7-point BARS format, with the number of dimensions ranging from five to nine, depending on the MOS. We computed a single composite score by averaging ratings across the individual scales for each MOS.

Early IOT&E evaluations noted low interrater reliability (IRR) estimates for the PRS (Moriarty & Bynum, 2011). Accordingly, we made several changes to the IMT instruments in an effort to improve their psychometric characteristics. First, we reduced the number of scales for the Army-wide PRS from eight to five, paralleling the five scores that had been generated from the original scales (Sparks & Peddie, 2013). To do this, we combined Effort and Discipline into a single dimension and combined what had been Peer Leadership and Support for Peers into a single “Working Effectively with Others” dimension. No changes were made to the MOS-specific PRS scale dimensions.

Second, we changed the rating scales for both the Army-wide and MOS-specific PRS from a 7-point BARS to a 5-point relative scale format with scales ranging from 1 (Among the Weakest) to 5 (Among the Strongest). The thought here was that a relative rating scale would more effectively mimic how cadre actually think about Soldiers in each class of trainees that they see for only limited periods of time.

The final IMT Army-wide rating scale dimensions and their associated definitions are shown in Table 6.1. A sample rating scale is provided in Figure 6.1.

Table 6.1. IMT PRS-S Army-Wide Rating Scale Dimensions and Associated Definitions

Scale	Definition
Effort and Discipline	• Puts forth individual effort in study, practice, preparation, and other training activities. • Behaves consistently with Army Core Values. • Shows respect in word and actions towards superiors, instructors, and others. • Obeys instructions, rules, and regulations.
Working Effectively with Others	• Respects, assists, cooperates, and teams with fellow Soldiers regardless of gender, race, ethnicity, ability, or background differences. • Supports other Soldiers. • When assigned as a leader, gains the cooperation of peers.
Physical Fitness and Bearing	• Meets basic Army APFT event standards. • Has sufficient physical endurance to complete most tasks. • Maintains self, uniforms, living areas, and barracks to Soldier standards.
Adjustment to the Army	• Demonstrates acceptable progress towards the completion of the Soldierization process. • Accepts and adheres to Army processes and requirements.
MOS Qualification Knowledge and Skill	• Learns and demonstrates AIT/OSUT knowledge and skills required for MOS qualification.
<i>Overall Performance</i>	Overall effectiveness

A. Effort and Discipline – Puts forth individual effort in study, practice, preparation, and other training activities. – Behaves consistently with Army Core Values. – Shows respect in word and actions towards superiors, instructors, and others. – Obeys instructions, rules, and regulations.				
1	2	3	4	5
Among the Weakest (in the bottom 20% of Soldiers)	Below Average (in the bottom 40% of Soldiers)	Average (better than the bottom 40% of Soldiers, but not as good as the top 40%)	Above Average (in the top 40% of Soldiers)	Among the Strongest (in the top 20% of Soldiers)

Figure 6.1. Sample IMT PRS-S rating scale.

Although the revised PRS appeared to nearly double IRR estimates for some scales, estimates were still low, ranging from .18 for Physical Fitness and Bearing to .37 for Effort and Personal Discipline (Kiger, Caramagno, & Reeder, 2017). The overall low estimates were likely due to multiple factors including rater fatigue, inability of rater to accurately differentiate among Soldiers, and the unproctored nature of the rating process. Moreover, previous research has found relatively low IRR estimates for military versus civilian samples (Van Iddekinge, Roth, Putka, & Lanivich, 2011). Additionally, the low IRRs may reflect raters viewing different samples of Soldiers' performance, and as such some of the low consistency between-raters may reflect true differences in observed performance (e.g., Putka, Hoffman, & Carter, 2014).

Data Cleaning

A Soldier's PRS-S 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 on either the original or revised scales), (b) omitted more than 10% of the assessment items, (c) indicated that he or she had not observed the Soldier on more than 50% of the dimensions, or (d) engaged in "flat responding"—that is, if the rater rated 10 or more Soldiers on a particular scale and 90% or more of those rating profiles were exactly the same. IMT PRS-S results are based on data from both the initial and revised PRS and are expressed on a 5-point scale.¹⁰

Interrater Reliability Estimates

Although most IMT and in-unit Soldiers were evaluated by a single rater on the PRS-S, there were at least two cadre raters for enough IMT Soldiers to compute IRR estimates (see Table 6.2). These estimates were disappointingly low, ranging from .20 to .28 across the six Army-wide scales.

Table 6.2. IMT PRS-S Interrater Reliability Estimates

	$G(q,k)$	k	n_{Ratees}	n_{Raters}
<i>Army-Wide</i>				
Adjustment to the Army	.21	1.05	9,121	805
Effort and Discipline	.25	1.05	9,250	823
MOS Qualification Knowledge & Skill	.21	1.06	8,590	784
Physical Fitness & Bearing	.26	1.04	8,718	760
Working Effectively with Others	.20	1.05	9,202	821
Overall Performance Rating	.28	1.05	9,178	817
<i>MOS-Specific</i>				
Infantryman (11B/C/X + 18X)	.23	1.03	1,134	88
Combat Engineer (12B)	.21	1.02	297	23
Fire Support Specialist (13F)	.22	1.02	145	14
Cavalry Scout (19D)	.09	1.01	147	9
Armor Crewman (19K)	.08	1.01	248	23
Military Police (31B)	.31	1.01	1,058	74
Combat Medic Specialist (68W)	.10	1.06	973	97
Motor Transport Operator (88M)	.11	1.09	1,128	150

Note. k = harmonic mean number of raters per ratee. Reliability for the PRS-S 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, as were data from the 7-point scale PRS-S administered prior to 2011.

¹⁰ The initial rating scale was converted from a 7-point scale to a 5-point scale by identifying meaningful cuts along the 7-point scale and comparing percentiles of the initial PRS-S to the revised PRS-S to ensure the cuts points produced consistent percentiles in each group. The following conversions were used: 1.00-2.99 = 1; 3.00-4.99 = 2; 5.00-5.99 = 3; 6.00-6.99 = 4; 7.00 = 5.

Descriptive Statistics

Table 6.3 summarizes the descriptive statistics for the IMT PRS-S data collected in the TOPS IOT&E. Mean ratings were above the mid-point, a consistent finding in prior Army research involving performance ratings (e.g., Campbell & Knapp, 2001; Knapp & Tremble, 2007; Moriarty & Bynum, 2011). Intercorrelations among the scales are provided in Appendix C (Table C.2).

The internal consistency (alpha) estimate for the Overall Performance composite rating was .93.

Table 6.3. Descriptive Statistics for the Supervisor Performance Rating Scales (PRS-S) in the IMT Validation Sample

Domain/PRS	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>
<i>Army-Wide</i>					
Overall Performance Composite	13,270	3.42	0.75	1.00	5.00
Adjustment to the Army	14,009	3.46	0.93	1.00	5.00
Effort and Discipline	14,146	3.35	0.96	1.00	5.00
MOS Qualification Knowledge & Skill	13,533	3.43	0.89	1.00	5.00
Physical Fitness & Bearing	13,629	3.37	0.93	1.00	5.00
Working Effectively with Others	14,097	3.39	0.94	1.00	5.00
Overall Performance Rating	14,084	3.51	0.85	1.00	5.00
<i>MOS-Specific</i>					
Infantryman (11B/C/X + 18X)	2,373	3.30	0.81	1.00	5.00
Combat Engineer (12B)	411	3.15	0.70	1.17	5.00
Fire Support Specialist (13F)	188	3.54	0.63	1.33	5.00
Cavalry Scout (19D)	315	2.97	0.56	1.13	4.75
Armor Crewman (19K)	420	3.36	0.71	1.00	5.00
Military Police (31B)	1,771	3.32	0.68	1.00	5.00
Human Resources Specialist (42A)	390	3.71	0.68	2.00	5.00
Combat Medic Specialist (68W)	1,615	3.22	0.82	1.00	5.00
Motor Transport Operator (88M)	1,842	3.65	0.69	1.20	5.00
All MOS Combined	9,458	3.37	0.77	1.00	5.00

Note. Ratings on PRS range from 1 and 5. PRS ratings from supervisors with a familiarity rating of 1 (“I have had little opportunity to observe this Soldier”) were excluded from analyses. Results based on fewer than 100 cases are not reported.

In-Unit PRS-S

Development

We based the in-unit Army-wide rating dimensions on a comprehensive set of performance dimensions identified in the Select21 research program (Knapp, Sager, & Tremble., 2005). Although the Army Class program was more recent, it used an abbreviated set of Army-wide dimensions because of its focus on Soldier classification. The TOPS project team conducted an in-house retranslation task to evaluate both the dimension definitions and behavioral statements that appeared on the Select21 PRS. The team gave particular attention to several scales that had shown exceptionally low interrater reliability in the Select21 research. This activity resulted in

13 dimensions that were used on the original in-unit TOPS PRS-S, in addition to an overall Leadership Potential scale (see Table 6.4). One scale with poor psychometric properties (involving interactions with indigenous people) was replaced in 2011 with the Adjustment to Army Life scale, comparable to a corresponding IMT scale.

Table 6.4. 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
Leadership Potential	Potential effectiveness as an NCO

As with the IMT PRS-S, the in-unit MOS-specific scales were either updated or developed in conjunction with the JKT development and review workshops described in Chapter 5. As mentioned previously, however, administration of MOS-specific in-unit PRS was discontinued early in the IOT&E so they are not described further here.

The in-unit PRS-S consistently employed the 7-point BARS format that was used for the original IMT scales (see Figure 6.2). We averaged ratings on the individual scales to form a single Overall Performance Composite score.

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 6.2. Sample in-unit supervisor rater performance rating scale.

Descriptive Statistics

Data cleaning protocols were the same as those described for the IMT PRS-P.

Table 6.5 reports basic descriptive statistics for the in-unit Army-wide PRS by performance domain. As with the IMT ratings, the means on all scales were well above the mid-point of the scale (3.5), though the full range of the 7-point scale was used at least by some raters. Intercorrelations among the scales are provided in Appendix C (Table C.3).

Table 6.5. Descriptive Statistics for the Supervisor Army-Wide Performance Rating Scales (PRS-S) in the In-Unit Validation Sample

PRS Dimensions/Composites	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>
Overall Performance Composite	6,354	5.11	1.15	1.00	7.00
Performing Core Warrior Tasks	6,291	4.84	1.41	1.00	7.00
Performing MOS-Specific Tasks	6,263	4.89	1.48	1.00	7.00
Communicating with Others	6,378	4.92	1.45	1.00	7.00
Processing Information	6,354	4.87	1.43	1.00	7.00
Solving Problems	6,270	4.80	1.41	1.00	7.00
Exhibiting Effort	6,310	4.93	1.51	1.00	7.00
Exhibiting Personal Discipline	6,371	5.26	1.48	1.00	7.00
Contributing to the Team	6,384	5.48	1.35	1.00	7.00
Exhibiting Fitness and Bearing	6,399	5.17	1.55	1.00	7.00
Adjusting to the Army	6,086	5.11	1.55	1.00	7.00
Following Safety Procedures	6,251	5.39	1.24	1.00	7.00
Developing Own Skills	6,244	4.85	1.41	1.00	7.00
Managing Personal Matters	6,099	5.42	1.44	1.00	7.00
Leadership Potential	6,269	4.64	1.67	1.00	7.00

Note. Ratings on PRS range from 1 to 7. PRS ratings from supervisors with a familiarity rating of 1 (“I have had little opportunity to observe this Soldier”) were excluded from analyses.

The majority of Soldiers in units were rated by only one supervisor, so we could not compute interrater reliability estimates. The internal consistency (alpha) estimate for the Overall Composite rating was .92.

Peer Performance Ratings

Background

In 2016, we began to develop scales and administration protocols to collect peer performance ratings. These efforts were primarily aimed at alleviating several issues with the collection of PRS-S data, particularly at IMT. To collect and match ratings for Soldiers in IMT, cadre visited an online system to retrieve unique Soldier identification numbers for each Soldier, then navigated to a second system to provide ratings for all the Soldiers in their class. This process required a significant time commitment from the cadre raters, due both to the multi-step process they had to follow as well as the large number of ratings they often had to make. The IMT peer performance rating system was designed to remove the need for cadre to serve as raters, as well as to take advantage of the close daily interactions that peers have with each other during training.

While developing the administration methods and ratings scales, we took into account the (a) ability level of initial entry Soldiers, (b) differences between the unproctored rating setting in IMT and the proctored setting in-unit, (c) potential rating biases that may arise when rating peers, and (d) the unique perspective peer raters could provide to evaluate certain aspects of performance.

Development

IMT PRS-P

Because of the especially challenging logistical difficulties associated with collecting IMT PRS-S data, we initially focused on development of peer performance ratings scales (PRS-P) for IMT. Early drafts of the IMT PRS-P duplicated the IMT PRS-S. We iteratively revised the scales 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 the Army Life Questionnaire (ALQ; see Chapter 5), we added 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. We also considered scales related to leadership, information processing ability, and problem-solving, but these dimensions were deemed more appropriate for the in-unit PRS-P. For each scale, we drafted example behaviors, some of which were taken from the PRS-S and some of which were based on the literature review conducted for the ALQ revisions. We also revised the response options so that the raters were comparing their rated peers to other Soldiers in the class. Finally, we removed the percentage benchmarks from the response options to improve interpretability for IMT raters.

The draft items were reviewed and revised internally to ensure that the scale labels were accurate and reflected the performance dimensions. We also made revisions to the example behaviors so that they were reflective of the rating scale and possible to enact in IMT, and we checked that the response option wording was appropriate for each rating dimension. After internal review, we

retained seven rating dimensions, not including overall performance, and 22 example behaviors. To provide a check on the scales, we conducted a retranslation exercise. Four psychologists who were not involved in the creation of the peer rating scales sorted the example behaviors into the performance scales. Eighteen of the 22 behavior examples were sorted into the expected performance scale, and revisions were made as needed to address the remaining behavior examples. There was no opportunity for Army SME review or pilot testing of the PRS-P; however, the components comprising the PRS-P (i.e., wording of response options and example behavioral statements) were reviewed by Army SMEs as part of the ALQ revisions.

The final IMT PRS-P scales are presented in Table 6.6. The rating scale anchors were all on a 1-5 scale, but varied by dimension. Figure 6.3 presents two examples. The IMT PRS-P includes the same familiarity question as the PRS-S, and each scale had a “Not Observed/Cannot Rate” response option.

Table 6.6. 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.
<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.

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 6.3. Sample IMT peer rater performance rating scales.

In-Unit PRS-P

After the IMT PRS-P was finalized, we turned to creating a version of the peer ratings scales for in-unit administration. We sought to align the in-unit PRS-P with the IMT PRS-P while making the scales appropriate for the in-unit context. We began the development of the in-unit PRS-P by reviewing each IMT PRS-P scale and editing the example behaviors to adapt them to the in-unit context. For example, in the Effort and Discipline scale, the example behavior “maintains self, uniforms, living areas, and barracks to Army standards” was edited to “Maintains self, uniforms, and work space to Army standards.” We also added new example behaviors that may not be enacted in the IMT context. For example, in the Counterproductive Soldier Behaviors scale, we added the example behavior “Speaks poorly about the Army to others.” As previously mentioned, we also added new scales to tap two additional dimensions, “Processing Information and Solving Problems” and “Leadership Potential.”

We iteratively edited the scales based on internal discussions, producing a draft instrument with nine scales, not including overall performance, and 30 example behaviors. We also conducted a second retranslation exercise. Four psychologists who were not involved in the scale development and who did not complete the IMT PRS-P retranslation exercise sorted the example behaviors into each scale. Twenty of the example behaviors were sorted into the expected scale. The remaining behaviors were edited as needed to increase their relevance to the scale or replaced with a more appropriate example. Again, it was not feasible to pilot test the scales prior to integration into the TOPS data collection activity.

The final in-unit PRS-P rating dimensions and example behaviors are presented in Table 6.7. As with the IMT scales, the rating scale anchors were all on a 1-5 scale, and the response options varied somewhat by dimension. With minor exceptions, the dimensions included on both the IMT and in-unit measures shared the same behavioral examples. Similarly, the in-unit PRS-P included the same familiarity question and “Not Observed/Cannot Rate” response option as the other PRS.

Table 6.7. 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 ^a	• 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
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.

^aOriginally titled *Leadership Potential*.

Administration

In 2017, we stopped collecting PRS-S data from IMT cadre and continued only collecting PRS-P data. The IMT cadre provided a copy of the class roster to Soldiers. After completing the other measures (e.g., JKT, ALQ), Soldiers were directed by instructional screens to find their name on the roster, then rate the five Soldiers who followed them on the roster. Use of the rosters significantly reduced the burden on cadre, while also providing the Soldiers with the correct spelling of the rated peers' names. This also meant, however, that rater-ratee data could only be matched based on names and this was an imperfect process. Of 82,314 IMT peer ratings collected, 69,018 (83.8%) were matched to a ratee.

To identify peer rater-ratee pairs in units, data collectors asked supervisors arriving to the test site to organize their Soldiers into groups according to who each supervisor should rate. As each group signed in, data collectors asked the Soldiers, as a group, if they had worked with each other long enough to be familiar with each other's performance. If the group included more than six Soldiers, only five peers were assigned to each Soldier. After all Soldiers and supervisors were signed in, data collectors completed a peer rating instruction sheet for each Soldier. This sheet listed the names and project identification numbers of the Soldiers for whom they were to provide ratings. Soldiers who were determined to have no peers were given a separate instruction sheet that contained instructions to skip the peer ratings section of the assessment. Of 2,920 in-unit peer ratings collected, 2,885 (98.8%) were matched to a ratee. Evidently, proctoring and the use of project identification numbers in addition to names helped maximize the match rate.

Data Cleaning

Data cleaning procedures for the PRS-P data paralleled that used for the PRS-S data. Our goal was to collect data from five peer raters for each Soldier. Table 6.8 provides information about the number of peer raters actually obtained for IMT and in-unit Soldiers after data cleaning. About 22% of IMT Soldiers and 31% of in-unit Soldiers had no peer ratings at all. It was clearly easier to obtain more than one peer rater at IMT than in-unit, but the fact that we obtained ratings from at least two peer raters for about 63% of the IMT Soldiers and 45% of the in-unit Soldiers is encouraging.

Table 6.8. Percentage of Soldiers with Various Numbers of Peer Raters

		Missing	1	2	3	4	5	6	more than 6
EOT	N	4480	2928	2956	3182	3082	2179	794	586
	%	22.19%	14.50%	14.64%	15.76%	15.27%	10.79%	3.93%	2.90%
IU	N	336	255	191	135	99	50	8	1
	%	31.26%	23.72%	17.77%	12.56%	9.21%	4.65%	0.74%	0.09%

PRS-P Interrater Reliability Estimates

Table 6.9 presents the IRR estimates for both 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.

Of note when reviewing these results is the Counterproductive Soldier Behaviors scale. This scale is reverse-scored and, although raters were alerted to the fact that the anchors were not the same across all dimensions, it is likely that this was a factor in the especially low interrater reliability estimates for that scale.

Table 6.9. PRS-P Interrater Reliability Estimates

IMT	$G(q,k)$	k	n_{Ratees}	n_{Raters}
Effort and Discipline	.46	2.16	8,969	8,238
Working Effectively with Other Soldiers	.41	2.18	8,987	8,268
Physical Fitness	.60	2.14	8,959	8,199
MOS Qualification Knowledge & Skill	.45	2.15	8,971	8,240
Resilience and Adjustment	.42	2.14	8,972	8,209
Counterproductive Soldier Behaviors	.25	2.10	8,909	7,990
Going Above and Beyond	.39	2.14	8,954	8,178
Overall Performance Rating	.50	2.18	8,984	8,255
In-Unit	$G(q,k)$	k	n_{Ratees}	n_{Raters}
Effort and Discipline	.31	1.65	409	357
Working Effectively with Other Soldiers	.20	1.67	410	361
Physical Fitness	.49	1.66	406	354
MOS Qualification Knowledge & Skill	.28	1.62	402	351
Processing Information & Solving Problems	.27	1.62	401	347
Resilience and Adjustment	.29	1.62	404	347
Counterproductive Soldier Behaviors	.09	1.63	406	350
Going Above and Beyond	.28	1.64	403	352
Emergent Leadership	.34	1.61	409	355
Overall Performance Rating	.37	1.66	409	357

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). Reliability for the PRS-P composite reflects coefficient Alpha. PRS ratings from peers with a familiarity rating of 1 (“I have had little opportunity to observe this Soldier”) were excluded from analyses.

PRS-P Scoring

Most of the scale-level ratings, particularly for the in-unit Soldiers, were too unreliably rated to consider using individually as criterion scores. We thus began our development of composite scores with a review of the scales. Eight of the 10 scales were administered at IMT (Information Processing and Leadership were excluded), and all were administered in-unit. Based on our review of the content, we hypothesized that the scales could potentially combine into multiple meaningful criterion composites.

Following this qualitative review, we conducted an exploratory factor analysis to examine the factor structure. We conducted these analyses separately for IMT and in-unit scales. In both cases, results of these analyses suggested a one-factor solution fit best (i.e., a single composite of all the scales).

As a next step, we examined whether a confirmatory factor analysis (CFA) also would support a single overall composite by comparing one- and two-factor models of the PRS-P scales. Conceptually, we divided the scales into “in-role” and “extra-role” dimensions for the two-factor model. Effort and Discipline, Physical Fitness and Bearing, MOS Qualification Knowledge and Skill, Information Processing, and Resilience and Adjustment were loaded onto the in-role factor; and Working with Others, Counterproductive Soldier Behaviors, Going Above and Beyond, and Leadership Potential were loaded onto the extra-role factor. The Information Processing and Leadership Potential scales were not included in the IMT model. All the available scales were loaded onto a single latent factor for the IMT and In-Unit one-factor models. Given the global nature of the Overall Performance rating, we excluded it from the CFA analyses.

Table 6.10 presents the model fit statistics for the IMT and In-Unit CFAs. Overall, each model fit the data well. Although the χ^2 difference test for the IMT model comparison indicated significantly better fit for the two-factor solution, the correlation between the in-role and extra role factors suggested that the two dimensions were practically identical ($r = .96$).

Table 6.10. CFA Model Fit Statistics for the Peer Performance Rating Scales (PRS-P)

Model	χ^2	df	CFI	RMSEA	SRMR	$\Delta\chi^2$
IMT						
One-factor	865.82	14	.97	.10	.03	
Two-factor	803.17	13	.97	.10	.03	62.65**
In-Unit						
One-factor	79.45	27	.97	.07	.03	
Two-factor	76.05	26	.97	.07	.03	3.41

** $p < .01$.

This analysis suggested that a single composite score would be the most suitable approach to capturing the peer performance ratings. Given the particularly low interrater reliability associated with the Counterproductive Soldier Behaviors scale, we excluded it from the overall composite score. We did, however, include the Overall Performance rating as we had for the IMT PRS-S overall composite score. Thus, we averaged scores across the scales (including Overall Performance and excluding Counterproductive Soldier Behaviors) to create the PRS-P Overall Performance criterion composite for both IMT and in-unit Soldiers.

PSR-P Descriptive Statistics

Tables 6.11 and 6.12 report basic descriptive statistics for the IMT and in-unit Army-wide PRS-P variables. As with the PRS-S ratings, the means on all scales were well above the mid-point of the scale (2.5), though the full range of the 7-point scale was used. Scale intercorrelations are provided in Appendix C (Table C.4). The internal consistency (alpha) estimate for the Overall Composite rating was .92 for both the IMT and in-unit PRS-P.

Table 6.11. 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	14,750	3.73	0.93	1	5
Working Effectively with Other Soldiers	14,773	3.74	0.92	1	5
Physical Fitness	14,741	3.67	0.89	1	5
MOS Qualification Knowledge & Skill	14,744	3.74	0.83	1	5
Resilience and Adjustment	14,748	3.67	0.87	1	5
Counterproductive Soldier Behaviors	14,638	3.80	1.01	1	5
Going Above & Beyond	14,731	3.49	0.94	1	5
Overall Performance Rating	14,772	3.73	0.87	1	5
Overall Performance Composite	14,784	3.68	0.76	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 6.12. 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	700	3.89	0.87	1	5
Working Effectively with Other Soldiers	703	4.02	0.84	1	5
Physical Fitness	692	3.71	0.95	1	5
MOS Qualification Knowledge & Skill	689	3.91	0.86	1	5
Information Processing	692	3.77	0.83	1	5
Resilience and Adjustment	695	3.87	0.85	1	5
Counterproductive Soldier Behaviors	695	3.85	1.07	1	5
Going Above & Beyond	692	3.57	0.95	1	5
Emergent Leadership	698	3.70	0.98	1	5
Overall Performance Rating	702	3.92	0.83	1	5
Overall Performance Composite	705	3.82	0.73	1.1	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.

Historical Summary of PRS Measures

Given the key role that performance ratings would ideally play in our criterion-related validation research, their psychometric properties have been disappointing. To help put the TOPS IOT&E experience in perspective and inform future efforts to improve ratings quality, we have consolidated information related to the use of Army-wide performance ratings in several major ARI research programs dating back to Project A (Campbell & Knapp, 2001). These include the following:

- Project A (1982-1993)
- NCO21 (2002-2004)

- Select21 (2004-2007)
- Army Class (2007-2010)
- TOPS (2010-2019)

With the exception of NCO21, which focused on validation of assessments for promotion purposes, these research programs involved the criterion-related validation of both operational and experimental pre-enlistment screening tests.

We have constructed an electronic library of Army-wide PRS materials from the five research programs to serve as a resource for future Army rating scale development efforts. The library includes (a) the rating scales, (b) administration protocols, and (c) excerpts from associated technical reports describing rating scale development, psychometric properties, and relations with ASVAB scores. Because interrater reliability cannot be computed with single raters and other predictors vary from one research project to the next, we hoped a strong sensible pattern of correlations with ASVAB scores would be an alternative strategy to gauge reliability and validity of the ratings. The remainder of this section provides a high-level synthesis of this information.

Rating Scale Development and Administration

Project A

Project A provided a luxury model example of criterion development and administration (Campbell & Knapp, 2001). The in-unit Army-wide and MOS-specific performance rating scales were developed through full-scale BARS procedures, including critical incident analysis to identify performance dimensions, development, scaling, and retranslation of rating scale anchors, and iterative rounds of pilot and field testing. The IMT scales were adapted from the in-unit scales. In-person proctors delivered extensive rater training, complete with visual aids of common rating errors, then looked over the shoulders of raters to point out instances of potential rating errors (e.g., halo). Supervisor (or instructor) and peer raters who indicated they had insufficient opportunity to work with a Soldier to make ratings were excused from the rating process. During data cleaning, data from raters indicating insufficient opportunity to observe or who provided ratings that clearly demonstrated suspect response patterns were excluded from the analysis database. In sum, Project A built a strong foundation in support of future ARI criterion-related research.

NCO21 and Select21

To develop the NCO21 and Select21 performance rating scales (which used the same format as the Project A scales), researchers created baseline sets of performance dimensions for first-term Soldiers and junior NCOs based on prior research, including Project A, Special Forces work, O*NET, and other sources (Ford, Campbell, Campbell, Knapp, & Walker, 2000). The performance dimensions and associated definitions were then iteratively reviewed and revised in a series of four workshops with roughly 40 senior NCOs in each (Knapp et al., 2002). Drafts of the rating scales themselves were derived from (a) applicable Project A scales, (b) applicable ARI Expanding the Concept of Quality in Personnel (EQUIP) project scales (Peterson et al., 1999), and (c) project staff using job analysis information. The rating scales were pilot tested and field tested before being finalized. The rating scale format was similar to that used in Project A, but included

low/moderate/high labels along the 7-point scale. The data collection and cleaning procedures were similar as well. Whereas Project A collected data from both supervisor/instructor and peer raters (in-unit and IMT), only in-unit supervisor data were collected in NCO21 and Select21.

Army Class

Army-wide IMT rating scales used in the Army Class longitudinal validation were derived from (a) critical incidents gathered in Select21, (b) Project A rating scales, and (c) Basic Combat Training rating scales that had been developed by ARI (Hoffman, Muraca, Heffner, Hendricks, & Hunter, 2009). Project staff drafted behavioral statements for each dimension that were retranslated and scaled by five senior NCOs. Instead of giving each dimension a single rating, participants rated multiple specific behaviors within each dimension, anchored by high and low-level variations of the behavioral statements on either end of the 5-point scales.

The Army Class in-unit scales and the MOS-specific rating scales (both IMT and in-unit) used the typical BARS format comparable to Project A and Select21. The applicable Project A and Select21 scales, where available, were used as a starting point. Performance dimensions and anchors were either eliminated or combined based on interrater reliability estimates obtained in prior research, with language updated as needed.

The Army Class research program tried an innovative method to improve the quality of the Army-wide IMT ratings, but there was limited involvement of SMEs in rating scale development and no opportunity for pilot or field testing. Moreover, Army Class was the first time Army rating data were collected via computer, eliminating in-person rater training and limiting the extent to which proctors could provide feedback to raters in the moment. Trying to minimize text and graphics to reduce raters' cognitive load also necessitated diminishing the extent of the rater training.

TOPS IOT&E

The TOPS IOT&E operated under similar, but even more pronounced constraints. Efforts to improve ratings quality during the course of the research included (a) shifting from an absolute to a relative rating scale for the IMT ratings, (b) dropping the in-unit MOS-specific ratings to reduce the rating task burden, (c) increasing the sensitivity and use of familiarity ratings to clean the ratings data, and (d) the introduction of peer ratings. The rating scales developed specifically for peer raters introduced important new assessment content and no doubt improved rating score reliability with the use of multiple raters. But these scales were created with limited degrees of traditional BARS-development procedures, SME review, and pilot testing. Shifting to peer ratings was a sensible action, but the content of the PRS-P may benefit from adjustment in the future. This might include eliminating the reverse-coded scale, sharpening conceptual distinctions among the dimensions being rated, and getting additional Army SME input into the scale language.

Patterns in Psychometric Qualities

Making comparisons across research projects (and variations within research projects) is complicated by many factors. Variables include rigor in development of the PRS instruments, administration format, administration conditions (proctoring, rater training, monitoring), Soldier

characteristics (e.g., trainees, first-term, second-term), number and type of raters, and data cleaning procedures.

Across all data collections, mean ratings were well above the scale mid-point. Interrater reliability estimates seemed to be highest in Project A. It is difficult to compare interrater reliability across research projects, however, because of differences in how these estimates were computed and the fact that several research projects did not have enough raters to estimate this value. There is some evidence that interrater reliability has suffered, perhaps in part, due to the lack of stand-up training and monitoring provided to raters that occurred when ratings started being administered via computer. It is also possible, however, that the mere fact of having several raters per Soldier (which was more typical of Project A than any of the other studies) led to better quality ratings.

We had hoped to gauge the impact of differences in how PRS were handled by examining patterns of relationships with ASVAB scores. As with the interrater reliability estimates, however, trends were difficult to discern. Some research projects reported correlations with AFQT and others with ASVAB factors and statistical corrections varied as well. Several research projects also reported predictor-criterion relationships using criterion scores that were composites of ratings and other measurement methods.

What do the Raters Say?

In 2017, TOPS data collectors interviewed 16 in-unit supervisor raters to get an idea of how they used the rating scales. Two supervisors admitted to skimming the rater instructions, but the others claimed to have read the instructions closely. Only one rater suggested that the instructions could be clearer. When asked if they found it difficult to make accurate ratings, most said no. Almost all of the supervisors said the rating scale behavioral examples were useful, but one indicated that they were not useful after making the first three or so ratings. Only one supervisor felt that it would have been helpful to have rating instructions provided by a proctor instead of online. Finally, the raters were shown a relative rating scale format similar to the one used to collect IMT ratings. The supervisors who responded were divided on this, with seven preferring the absolute rating scale and eight preferring the relative scale format.

Summary

As part of the TOPS IOT&E, we have collected performance ratings data from IMT cadre and in-unit supervisors, as well as peers of both IMT and in-unit Soldiers. High scale intercorrelations have resulted in using overall performance composite scores which do have high internal consistency. Where we have evidence of consistency across raters, however, this is weaker than desired. That said, the peer ratings which are generally based on multiple instead of single raters do yield higher inter-rater reliability estimates than the IMT cadre ratings.

As noted previously, shifting to peer ratings was a sensible action, but the content of the PRS-P may benefit from adjustment for future use. Processes for identifying raters (particularly in-units) and matching rater-ratee pairs (particularly in IMT) might also improve with lessons learned from this early experience.

As it had been for a series of predecessor ARI research programs, obtaining psychometrically sound performance ratings was a major goal in the TOPS IOT&E. Adjustments made before and during the course of the IOT&E met with minimal apparent improvement in data quality. This is not terribly surprising given that the general research literature as well as our own analysis of prior Army research suggests that rating scale formats make little difference in the psychometric quality of ratings (Pulakos, 2007). That said, future research should capitalize as much as possible from past efforts, while continuing the quest.

Some of those best practices include the following:

- Minimizing instructional and behavioral anchor text (without losing too much content) and maximizing the use of graphics to support instruction
- Keeping the number of ratings collected from each rater to a minimum, both in terms of number of dimensions to be rated and number of ratees
- Using a relatively small number of conceptually independent dimensions
- Pilot testing or otherwise engaging knowledgeable Army personnel in review and revision of the rating scales
- Ideally, using IRR estimates to drop scales with lowest reliabilities, so they do not detract from composite score reliability

In an unrelated effort, ARI is experimenting with development of Computer Adaptive Rating Scales (CARS). This research is in progress so it is too soon to tell how well it might work. Early development work did make it clear, however, that not all performance dimensions are suited to the CARS method, making it unlikely that it will fully satisfy future Army criterion-related research needs.

CHAPTER 7: EVIDENCE FOR THE PREDICTIVE VALIDITY OF THE TAPAS

Michael G. Hughes (HumRRO)

This chapter describes our evaluation of the Tailored Adaptive Personality Assessment System (TAPAS) and its potential 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 two different analytic approaches to conduct this evaluation with both initial military training (IMT) and in-unit criteria. First, we describe meta-analytic analyses in which we estimated corrected correlations between the TAPAS composites and Soldier outcomes. Next, we discuss regression analyses that examined the incremental validity of the TAPAS to predict outcomes beyond the Armed Forces Qualification Test (AFQT).

Across both approaches, we evaluated the 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., Army Life Questionnaire; ALQ) and performance ratings. Table 7.1 lists the IMT and in-unit criteria used in this chapter and provides a brief description of each. For composite criteria, the component scales are provided. Intercorrelations among the selected criteria are provided in Appendix C (Table C.5).

Table 7.1. Criteria Used to Examine the Validity of the TAPAS

Criterion	Associated Scales	Description
<i>IMT</i>		
PRS-P 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-Socionormative ALQ Organizational Citizenship Behavior/Leadership	General aptitude for effective peer leadership in the Army. Scales are averaged to form the composite.

Table 7.1. (Continued)

Criterion	Associated Scales	Description
<i>IMT</i>		
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 & Skill	WTBD JKT MOS JKT AIT Grade	WTBD JKT and MOS JKT are averaged to form an overall knowledge/skill composite. For those that do not have an MOS JKT score, AIT grade is substituted when available.
WTBD JKT	—	General Army knowledge that applies to all enlisted Soldiers.
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 if a Soldier restarted IMT one or more times.
<i>In-Unit</i>		
PRS-S 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.
PRS-P 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.

Table 7.1. (Continued)

Criterion	Associated Scales	Description
<i>In-Unit</i>		
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-Socionormative 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 & Skill	WTBD JKT MOS JKT AIT Grade	WTBD JKT and MOS JKT are averaged to form an overall knowledge/skill composite. For those that do not have an MOS JKT score, AIT grade is substituted when available.
WTBD JKT	—	General Army knowledge that applies to all enlisted Soldiers.
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.
<i>Attrition</i>		
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.

Note. For the Knowledge & Skill composite, only a small proportion of Soldiers have AIT grade data; therefore, the vast majority of Knowledge & Skill composite scores do not include AIT grades. Additionally, administration of In-Unit MOS-specific JKTs was discontinued in 2017. Most Knowledge & Skill composite scores since the discontinuation are equivalent to scores on the WTBD JKT.

Meta-Analytic Analyses

Multiple versions of the TAPAS have been administered since its first use by the Army. As such, data from the TAPAS reflect not one but many versions of the measure, each with some number of different items. Similarly, data from each version was collected from different samples of Soldiers and, in some cases, at different periods of time (see Table 3.2). Given these factors, we conducted a meta-analysis to account for the different versions and to provide more accurate estimates of the relationships each TAPAS composite has with key Soldier outcomes.

Meta-analysis is a quantitative method that seeks to cumulate research findings across different studies to provide a more accurate representation of a phenomena than any single study alone (Hunter & Schmidt, 1990; Hunter, Schmidt, & Jackson, 1982). In cumulating across studies, meta-analytic methods seek to remove statistical error (artifacts) in observed estimates. Although there are numerous sources of potential error, sampling error (which stems from a particular sample not being completely representative of an entire population) is the primary artifact addressed by meta-analysis.

For the present analysis, we conceptualized data associated with each TAPAS version as a separate “study” in predicting a given Soldier outcome. In total, the TOPS database contained data from nine TAPAS versions. Given the number of Soldiers with both TAPAS and criterion data, each version may have a large enough sample such that sampling error is likely minimal. Nonetheless, meta-analytically combining the observed correlations between each TAPAS version and outcome should yield a more accurate estimate of the relationships.

In addition to sampling error, we also accounted for imperfect measurement of Soldier outcomes (i.e., criterion unreliability). As such, we computed corrected estimates of the correlations between the TAPAS composites and outcomes. For job knowledge and self-report measures (e.g., ALQ), we used coefficient alpha as an estimate of reliability. For the PRS-P criterion composites, we used the average interrater reliability estimates across the component scales in the correction. For dichotomous outcomes, we applied a different correction due to artificially reduced effects associated with dichotomization (Hunter & Schmidt, 1990). Specifically, we applied Kemery, Dunlap, and Griffeth’s (1988) method that has been shown to produce accurate corrections to point-biserial correlations (Williams, 1990). This correction adjusts the correlation for proportions less than $p = .50$. Finally, we did not apply a correction for unreliability of the TAPAS given its history and potential future use as an operational screening tool (Drasgow, Whetzel, & Oppler, 2007).

In contrast to the subsequent incremental validity analyses, the meta-analytic analyses relied on zero-order correlations as is traditionally used by meta-analysis. That is, the effects of AFQT were not included in these analyses. However, the ability of the TAPAS to predict beyond the AFQT is demonstrated by the incremental validity analyses for a number of outcomes. See Appendix E for uncorrected bivariate correlations between scores on the TAPAS scales and the selected criterion measures.

Predicting IMT Performance

Table 7.2 shows results of the meta-analytic analyses of the TAPAS Can-Do, Will-Do, and Adaptation composites with IMT technical performance outcomes. As expected, the Can-Do

composite exhibited the strongest effects with the technical performance outcomes, although the correlations were generally small (Knowledge & Skill $\bar{r} = .22$; WTBD JKT $\rho = .24$). We did not compute the reliability of Knowledge & Skill and therefore did not estimate a correlation corrected for criterion unreliability. Knowledge & Skill includes different measures across Soldiers, including MOS-specific JKTs and AIT grades when available.

Table 7.2. Meta-analytic Results for TAPAS Correlations with IMT Technical Performance

Dependent Variable	<i>k</i>	<i>N</i>	$\bar{r}$	95% CI	ρ	σ_{ρ}
Can-Do ^a						
Knowledge & Skill ^b	6	31,346	.216	.199 — .234	—	—
WTBD JKT	6	44,851	.195	.178 — .212	.237	.022
Will-Do						
Knowledge & Skill ^b	8	37,655	.034	.019 — .050	—	—
WTBD JKT	8	53,507	.031	.025 — .037	.038	.000
Adaptation						
Knowledge & Skill ^b	8	37,655	.094	.081 — .107	—	—
WTBD JKT	8	53,507	.087	.077 — .096	.105	.007

Note. WTBD JKT = Warrior Tasks and Battle Drills Job Knowledge Test. *k* = number of TAPAS versions in the analysis. *N* = total number Soldiers across TAPAS versions. $\bar{r}$ = sample size-weighted mean correlation. 95% CI = estimated 95% confidence interval for the mean correlation. ρ = estimated true score correlation. σ_{ρ} = estimated true standard deviation for the correlation. ρ is corrected for criterion unreliability.

^a No data were available for the Can-Do TAPAS versions V7 and V8.

^b Reliability was not computed for the Knowledge & Skill composite because the composite reflects different measures/items across Soldiers. Thus, the correlation was not corrected for unreliability.

Table 7.3 presents meta-analytic results for ALQ scales, criterion composites, and APFT. As expected and in line with the incremental validity results reported later in this chapter, the Will-Do composite has the strongest relationships with these outcomes. Peer Leadership and Resilience exhibited the strongest of these corrected correlations (ρ s = .24 and .23, respectively). Reliability of the APFT was unknown and therefore we did not compute a corrected correlation. Nonetheless, the sample size-weighted correlation with the Will-Do composite was the largest among these outcomes ($\bar{r} = .25$). In general, outcomes with the strongest corrected correlations were those for which the associated TAPAS composites demonstrated the greatest incremental validity.

Table 7.4 presents the meta-analytic results for the PRS composites. For these analyses, we corrected the correlations using the mean interrater reliability of the component scales. As expected, corrected correlations were near-zero for the Can-Do composite. Small corrected correlations were found for the Will-Do and Adaptation composites. Of these, the strongest relationships were between Will-Do and both the PRS-S and PRS-P Overall Performance Composites (ρ s = .25 and .22, respectively).

Table 7.3. Meta-analytic Results for TAPAS Correlations with IMT ALQ and APFT Criteria

Dependent Variable	<i>k</i>	<i>N</i>	$\bar{r}$	95% CI	ρ	σ_{ρ}
Can-Do ^a						
Commitment & Fit	6	51,349	.022	.001 — .036	.024	.015
Retention Cognitions	6	51,127	-.009	-.020 — .002	-.010	.010
Peer Leadership	6	16,465	.038	.032 — .045	.042	.000
Counterproductive Behavior	6	16,465	-.020	-.030 — -.009	-.021	.000
Resilience	6	16,463	.028	.013 — .043	.031	.000
APFT Score ^b	6	49,393	.016	.006 — .027	—	—
Will-Do						
Commitment & Fit	8	60,392	.121	.116 — .126	.134	.000
Retention Cognitions	8	60,134	.045	.040 — .050	.047	.000
Peer Leadership	8	16,620	.216	.189 — .243	.240	.036
Counterproductive Behavior	8	16,620	-.078	-.099 — -.058	-.085	.022
Resilience	8	16,618	.207	.182 — .231	.230	.032
APFT Score ^b	8	58,132	.248	.236 — .259	—	—
Adaptation						
Commitment & Fit	8	60,392	.020	.012 — .028	.023	.003
Retention Cognitions	8	60,134	-.031	-.037 — -.024	-.032	.000
Peer Leadership	8	16,620	-.018	-.030 — -.005	-.020	.000
Counterproductive Behavior	8	16,620	-.004	-.023 — .015	-.004	.017
Resilience	8	16,618	.078	.062 — .095	.087	.011
APFT Score ^b	8	58,132	.168	.159 — .178	—	—

Note. *k* = number of TAPAS versions in the analysis. *N* = total number Soldiers across TAPAS versions. $\bar{r}$ = sample size-weighted mean correlation. 95% CI = estimated 95% confidence interval for the mean correlation. ρ = estimated true score correlation. σ_{ρ} = estimated true standard deviation for the correlation. ρ is corrected for criterion unreliability.

^a No data were available for the Can-Do TAPAS versions V7 and V8.

^b Reliability was not available for the APFT. Thus, the correlation was not corrected for unreliability.

Table 7.4. Meta-analytic Results for TAPAS Correlations with IMT Performance Rating Criteria

Dependent Variable	<i>k</i>	<i>N</i>	$\bar{r}$	95% CI	ρ	σ_{ρ}
Can-Do ^a						
PRS-S: Overall Performance Composite	6	10,395	.022	.013 — .032	.050	.000
PRS-P: Overall Performance Composite ^b	5	14,307	.023	.015 — .030	.035	.000
Will-Do						
PRS-S: Overall Performance Composite	8	12,559	.109	.089 — .129	.246	.031
PRS-P: Overall Performance Composite ^b	7	14,426	.145	.130 — .159	.219	.000
Adaptation						
PRS-S: Overall Performance Composite	8	12,559	.068	.050 — .086	.149	.000
PRS-P: Overall Performance Composite ^b	7	14,426	.117	.109 — .126	.178	.000

Note. PRS = Performance Rating Scales, where -S denotes supervisor ratings and -P denotes peer ratings. *k* = number of TAPAS versions in the analysis. *N* = total number Soldiers across TAPAS versions. $\bar{r}$ = sample size-weighted mean correlation. 95% CI = estimated 95% confidence interval for the mean correlation. ρ = estimated true score correlation. σ_{ρ} = estimated true standard deviation for the correlation. ρ is corrected for criterion unreliability.

^a No data were available for the Can-Do TAPAS versions V7 and V8.

^b TAPAS versions were excluded from the analysis where *n* < 5.

Predicting In-Unit Performance

The in-unit meta-analytic findings demonstrate similarities with those from the IMT outcomes. As shown in Table 7.5, the Can-Do TAPAS composite had a moderate corrected correlation with WTBD JKT ($\rho = .32$). For both technical performance outcomes, Can-Do exhibited positive correlations despite *not* providing incremental gains in prediction beyond the AFQT (as observed from the incremental validity analyses later in the chapter). As previously stated, the AFQT is a strong predictor of such performance. Thus, the TAPAS does not provide additional explanatory value for these criteria regardless of their relationships with the Can-Do composite.

Table 7.5. Meta-analytic Results for TAPAS Correlations with In-Unit Technical Performance

Dependent Variable	<i>k</i>	<i>N</i>	$\bar{r}$	95% CI	ρ	σ_ρ
Can-Do ^a						
Knowledge & Skill ^b	6	4,523	.241	.194 — .287	—	—
WTBD JKT	6	4,517	.243	.200 — .286	.324	.056
Will-Do						
Knowledge & Skill ^b	8	6,684	.022	-.013 — .057	—	—
WTBD JKT	8	6,677	.027	-.001 — .055	.036	.027
Adaptation						
Knowledge & Skill ^b	8	6,684	.099	.064 — .133	—	—
WTBD JKT	8	6,677	.108	.079 — .137	.145	.032

Note. WTBD JKT = Warrior Tasks and Battle Drills Job Knowledge Test. *k* = number of TAPAS versions in the analysis. *N* = total number Soldiers across TAPAS versions. $\bar{r}$ = sample size-weighted mean correlation. 95% CI = estimated 95% confidence interval for the mean correlation. ρ = estimated true score correlation. σ_ρ = estimated true standard deviation for the correlation. ρ is corrected for criterion unreliability.

^a No data were available for the Can-Do TAPAS versions V7 and V8.

^b Reliability was not computed for the Knowledge & Skill composite because the composite reflects different measures/items across Soldiers. Thus, the correlation was not corrected for unreliability.

Table 7.6 shows small corrected correlations between the Will-Do TAPAS composite and both In-Unit Peer Leadership and Resilience ($.20 < \rho \leq .24$). Additionally, both Will-Do and Adaptation showed small positive correlations with the APFT. Overall, these results resemble the meta-analytic findings with the IMT ALQ and APFT criteria.

Table 7.7 shows that the TAPAS is not strongly related to the in-unit PRS in general. Of the three TAPAS composites, Will-Do exhibited the strongest corrected correlation with the PRS-P Overall Performance composite ($\rho = .18$). Note that because interrater reliability is unavailable for the PRS-S in-unit measures, we did not apply a correction.

Table 7.6. Meta-analytic Results for TAPAS Correlations with In-Unit ALQ and APFT Criteria

Dependent Variable	<i>k</i>	<i>N</i>	$\bar{r}$	95% CI	ρ	σ_ρ
Can-Do ^a						
Commitment & Fit	6	5,499	.005	-.017 — .027	.005	.000
Retention Cognitions ^b	5	839	.003	-.059 — .064	.003	.000
Peer Leadership ^b	5	840	.034	-.051 — .118	.037	.064
Counterproductive Behavior	5	840	.030	-.009 — .069	.034	.000
Resilience	5	840	.026	-.009 — .061	.029	.000
APFT Score ^c	6	5,236	-.018	-.031 — -.004	—	—
Will-Do						
Commitment & Fit	8	7,857	.065	.048 — .083	.072	.000
Retention Cognitions ^b	7	998	.016	-.025 — .056	.016	.000
Peer Leadership ^b	7	999	.217	.155 — .279	.240	.026
Counterproductive Behavior	7	999	-.052	-.127 — .024	-.058	.065
Resilience	7	999	.180	.132 — .228	.202	.000
APFT Score ^c	8	7,516	.217	.201 — .234	—	—
Adaptation						
Commitment & Fit	8	7,857	.028	.017 — .040	.031	.000
Retention Cognitions ^b	7	998	.051	.001 — .101	.053	.000
Peer Leadership ^b	7	999	.043	-.022 — .107	.047	.026
Counterproductive Behavior	7	999	.046	.014 — .079	.052	.000
Resilience	7	999	.101	.037 — .166	.114	.027
APFT Score ^c	8	7,516	.145	.118 — .172	—	—

Note. *k* = number of TAPAS versions in the analysis. *N* = total number Soldiers across TAPAS versions. $\bar{r}$ = sample size-weighted mean correlation. 95% CI = estimated 95% confidence interval for the mean correlation. ρ = estimated true score correlation. σ_ρ = estimated true standard deviation for the correlation. ρ is corrected for criterion unreliability.

^a No data were available for the Can-Do TAPAS versions V7 and V8.

^b TAPAS versions were excluded from the analysis where *n* < 5.

^c Reliability was not available for the APFT. Thus, the correlation was not corrected for unreliability.

Table 7.7. Meta-analytic Results for TAPAS Correlations with In-Unit Performance Rating Criteria

Dependent Variable	<i>k</i>	<i>N</i>	$\bar{r}$	95% CI	ρ	σ_ρ
Can-Do ^a						
PRS-S: Overall Performance Composite	6	4,161	.042	.034 — .050	—	—
PRS-P: Overall Performance Composite ^b	5	614	-.043	-.144 — .058	-.077	.126
Will-Do						
PRS-S: Overall Performance Composite	8	6,016	.103	.092 — .113	—	—
PRS-P: Overall Performance Composite ^b	7	691	.102	.027 — .176	.181	.009
Adaptation						
PRS-S: Overall Performance Composite	8	6,016	.064	.054 — .074	—	—
PRS-P: Overall Performance Composite ^b	7	691	.016	-.060 — .091	.028	.022

Note. *k* = number of TAPAS versions in the analysis. *N* = total number Soldiers across TAPAS versions. $\bar{r}$ = sample size-weighted mean correlation. 95% CI = estimated 95% confidence interval for the mean correlation. ρ = estimated true score correlation. σ_ρ = estimated true standard deviation for the correlation. ρ is corrected for criterion unreliability.

^a No data were available for the Can-Do TAPAS versions V7 and V8.

^b TAPAS versions were excluded from the analysis where *n* < 5.

Predicting Dichotomous Outcomes

Results of the meta-analyses examining the relationships between the TAPAS composites and dichotomous outcomes are presented in Tables 7.8 for administrative outcomes and 7.9 for attrition. As previously stated, these correlations were adjusted to estimate the relationships given a standardized, higher base rate of the outcomes (Kemery et al., 1988). We applied this correction to eliminate differences in base rates across the TAPAS version samples and because both the administrative and attrition criteria have very low base rates ($ps < .26$ and $.22$, respectively). For the dichotomous administrative criteria, the TAPAS composites generally showed negligible relationships with the outcomes, with the exception of IMT Disciplinary Incidents, which had a small corrected point-biserial correlation with the Will-Do composite ($\rho = -.11$). Regarding the attrition outcomes, the Adaptation composite was the strongest of the predictors overall ($\rho s = -.09$ for all three time points). Additionally, the Will-Do composite had similar corrected point-biserial correlations for attrition at 6 and 12 months ($\rho s = -.09$).

Table 7.8. Meta-analytic Results for TAPAS Correlations with Administrative Criteria

Dependent Variable	<i>k</i>	<i>N</i>	$\bar{r}$	95% CI	ρ
Can-Do ^a					
IMT Disciplinary Incidents	6	49,620	-.016	-.024 — -.008	-.017
IMT Restarts	6	465,945	-.005	-.013 — .002	-.002
In-Unit Disciplinary Incidents	6	5,499	-.013	-.040 — .014	-.014
Will-Do					
IMT Disciplinary Incidents	8	58,652	-.104	-.113 — -.096	-.113
IMT Restarts	8	593,580	-.026	-.028 — -.025	-.037
In-Unit Disciplinary Incidents	8	7,857	-.061	-.086 — -.036	-.066
Adaptation					
IMT Disciplinary Incidents	8	58,652	-.062	-.074 — -.050	-.067
IMT Restarts	8	593,580	-.023	-.027 — -.020	-.039
In-Unit Disciplinary Incidents	8	7,857	-.038	-.059 — -.016	-.041

Note. *k* = number of TAPAS versions in the analysis. *N* = total number Soldiers across TAPAS versions. $\bar{r}$ = sample size-weighted mean correlation. 95% CI = estimated 95% confidence interval for the mean correlation. ρ = estimated true score correlation corrected for an optimal proportion of $p=.50$.

^a No data were available for the Can-Do TAPAS versions V7 and V8.

Table 7.9. Meta-analytic Results for TAPAS Correlations with Cumulative Attrition through 24 Months of Service (Regular Army Only)

Dependent Variable	<i>k</i>	<i>N</i>	$\bar{r}$	95% CI	ρ
Can-Do ^a					
6 Month	6	253,804	-.015	-.019 — -.010	-.020
12 Month	6	219,415	-.013	-.021 — -.006	-.016
24 Month	6	164,618	-.022	-.036 — -.008	-.025
Will-Do					
6 Month	8	320,593	-.066	-.069 — -.062	-.090
12 Month	8	285,695	-.071	-.074 — -.068	-.087
24 Month	8	229,487	-.065	-.068 — -.062	-.074
Adaptation					
6 Month	8	320,593	-.065	-.068 — -.062	-.089
12 Month	8	285,695	-.073	-.076 — -.070	-.089
24 Month	8	229,487	-.075	-.079 — -.070	-.086

Note. *k* = number of TAPAS versions in the analysis. *N* = total number Soldiers across TAPAS versions. $\bar{r}$ = sample size-weighted mean correlation. 95% CI = estimated 95% confidence interval for the mean correlation. ρ = estimated true score correlation corrected for an optimal proportion of $p=.50$.

^a No data were available for the Can-Do TAPAS versions V7 and V8.

Incremental Validity Analyses

The purpose of the incremental validity analyses was to examine the potential of the TAPAS composites to predict Soldier outcomes beyond the AFQT. Whereas the results of our meta-analytic approach provided true relationships between the TAPAS and key criteria, the incremental validity analyses studied the effects of the TAPAS while accounting for the effect of the AFQT. For each TAPAS composite and criterion of interest, we used hierarchical regression to examine the incremental validity of the TAPAS after accounting for the effect of the AFQT. This approach was generally consistent with previous evaluations of the TAPAS and similar experimental non-cognitive predictors (e.g., Ingerick, Diaz, & Putka, 2009; Knapp & Heffner, 2009; 2010; Trippe, Caramagno, Allen, & Ingerick, 2011).

We conducted the hierarchical regression analyses using two different modeling approaches. For our primary approach, 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 past operational use of the TAPAS. Results of these analyses are presented in Tables 7.10 to 7.17. For our second approach, we entered in Step 2 of the models the individual TAPAS scales that comprise each composite. This analysis provides a more theoretical (vs. operational) perspective of the relationships that the Can-Do, Will-Do, and Adaptation TAPAS *scales* have with key criteria.¹¹ The results of this secondary approach are presented in Figures 7.1 to 7.4.

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.,

¹¹ For the TAPAS, the Can-Do composite consists of six scales, and the Will-Do and Adaptation composites each consist of four scales.

attrition, Disciplinary Incidents¹², IMT Restarts). We discuss the results of the OLS regression analyses with respect to *R* values when interpreting the validity of the TAPAS. We refer primarily to odds ratios (ORs) when interpreting the results of the logistic regression analyses. Additional details concerning the specific indices presented for the logistic regression analyses are provided with the results of the dichotomous outcomes later in the chapter. Attrition criteria were examined for Regular Army Soldiers only.

Predicting IMT Performance

Tables 7.10 to 7.12 summarize the incremental validity results of the TAPAS composites for predicting IMT performance criteria over and above the AFQT. Overall, the results suggest that both the Will-Do and Adaptation TAPAS composites can enhance the Army's ability to predict a number of important outcomes. Below, we describe the specific outcomes for which the TAPAS composites demonstrated notable predictive gains beyond the AFQT alone.

With respect to the motivation-based performance criteria, the TAPAS composites exhibited gains in predictive validity over the AFQT in predicting several outcomes. The TAPAS Will-Do composite enhanced the prediction of the Resilience scale ($\Delta R = .19$) Peer Leadership composite ($\Delta R = .17$), and APFT Score ($\Delta R = .16$); and to a lesser extent, the prediction of Commitment and Fit ($\Delta R = .09$) and Counterproductive Soldier Behavior ($\Delta R = .06$). It also showed a slight enhancement in the prediction of both the supervisor and peer PRS Overall Performance composites (ΔR s = .05 and .10, respectively). These incremental gains are likely attenuated due to the low interrater reliability of these measures.

The TAPAS Adaptation composite also demonstrated incremental validity in the prediction of APFT Score ($\Delta R = .09$), Resilience ($\Delta R = .07$), and PRS-P: Overall Performance composite ($\Delta R = .07$). The Can-Do composite is not designed to predict Will Do criteria, and it did not demonstrate any notable incremental validity for the motivation-based outcomes.

Table 7.10. Incremental Validity Estimates for the TAPAS over AFQT for Predicting IMT Technical Performance

IMT Criterion Measure / Model	AFQT <i>R</i>	AFQT + TAPAS <i>R</i>	ΔR
<i>Knowledge & Skill</i>			
	<i>n</i> = 31,346 - 37,655		
Can-Do ^a	.43	.44	.00
Will-Do	.43	.43	.00
Adaptation ^a	.43	.43	.00
<i>WTBD JKT</i>			
	<i>n</i> = 44,851 - 53,507		
Can-Do ^a	.41	.41	.00
Will-Do	.41	.41	.00
Adaptation ^a	.41	.41	.00

Note. WTBD JKT = Warrior Tasks and Battle Drills Job Knowledge Test. *R* = multiple correlations between the AFQT and selected TAPAS composite scales with the targeted criterion measure. ΔR = Increment in *R* from adding the selected TAPAS composite scale to the regression model [(AFQT + TAPAS) – AFQT Only]. Bolded values indicate $p < .05$.

^a Because the Can-Do and Adaptation composites are based on a subset of the data, these results may underestimate their incremental validity.

¹² The dichotomous version of Disciplinary Incidents (0 = no disciplinary incidents; 1 = one or more disciplinary incidents) was used for all analyses due to a low base rate beyond one incident.

Table 7.11. Incremental Validity Estimates for the TAPAS over AFQT for Predicting IMT ALQ and APFT Criteria

IMT Criterion Measure / Model	AFQT R	AFQT + TAPAS R	ΔR
<i>Commitment & Fit</i> <i>n</i> = 51,349 - 60,392			
Can-Do ^a	.04	.05	.01
Will-Do	.04	.13	.09
Adaptation ^a	.04	.05	.01
<i>Retention Cognitions</i> <i>n</i> = 51,127 - 60,134			
Can-Do ^a	.11	.12	.00
Will-Do	.11	.12	.01
Adaptation ^a	.11	.11	.00
<i>Peer Leadership</i> <i>n</i> = 16,465 - 16,620			
Can-Do ^a	.05	.08	.03
Will-Do	.05	.23	.17
Adaptation ^a	.05	.05	.00
<i>Counterproductive Soldier Behavior</i> <i>n</i> = 16,465 - 16,620			
Can-Do ^a	.02	.02	.00
Will-Do	.02	.08	.06
Adaptation ^a	.02	.02	.00
<i>Resilience</i> <i>n</i> = 16,463 - 16,618			
Can-Do ^a	.02	.04	.02
Will-Do	.02	.21	.19
Adaptation ^a	.02	.08	.07
<i>APFT Score</i> <i>n</i> = 49,393 - 58,132			
Can-Do ^a	.09	.10	.00
Will-Do	.10	.26	.16
Adaptation ^a	.10	.18	.09

Note. R = multiple correlations between the AFQT and selected TAPAS composite scales with the targeted criterion measure. ΔR = Increment in R from adding the selected TAPAS composite scale to the regression model [(AFQT + TAPAS) – AFQT Only].

Bolded values indicate $p < .05$.

^a Because the Can-Do and Adaptation composites are based on a subset of the data, these results may underestimate their incremental validity.

Table 7.12. Incremental Validity Estimates for the TAPAS over AFQT for Predicting IMT Performance Rating Criteria

IMT Criterion Measure / Model	AFQT R	AFQT + TAPAS R	ΔR
<i>PRS-S: Overall Performance Composite</i> <i>n</i> = 10,395 - 12,559			
Can-Do ^a	.04	.04	.00
Will-Do	.06	.11	.05
Adaptation ^a	.06	.08	.02
<i>PRS-P: Overall Performance Composite</i> <i>n</i> = 14,311 - 14,430			
Can-Do ^a	.06	.06	.00
Will-Do	.05	.15	.10
Adaptation ^a	.05	.12	.07

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

^a Because the Can-Do and Adaptation composites are based on a subset of the data, these results may underestimate their incremental validity.

Regarding the cognitive criteria, the TAPAS composites (Can-Do, Will-Do, and Adaptation) evidenced no notable increments over the AFQT in predicting scores on the composite measure of Knowledge & Skill or WTBD JKT ($\Delta R_s < .01$). These results are not surprising given that the AFQT is an established strong predictor of Can Do outcomes. However, the Can-Do TAPAS composite is positively related to these outcomes as demonstrated by the meta-analytic results.

Figures 7.1 and 7.2 provide results of analyses examining the incremental validity of criteria-specific TAPAS scales over and above the AFQT for Will Do and Can Do criteria, respectively. Note that the results shown in these figures reflect regression models that used individual TAPAS scales as predictors (as opposed to the TAPAS composites used above). Similar to the results of the TAPAS composite models discussed previously, the Will-Do scales provided the largest gains over the AFQT. Increases in R were largest for the Peer Leadership, Resilience, and APFT Score ($\Delta R_s > .18$). Results of the Can-Do TAPAS scales showed no notable increments over the AFQT in predicting the Can Do criteria.

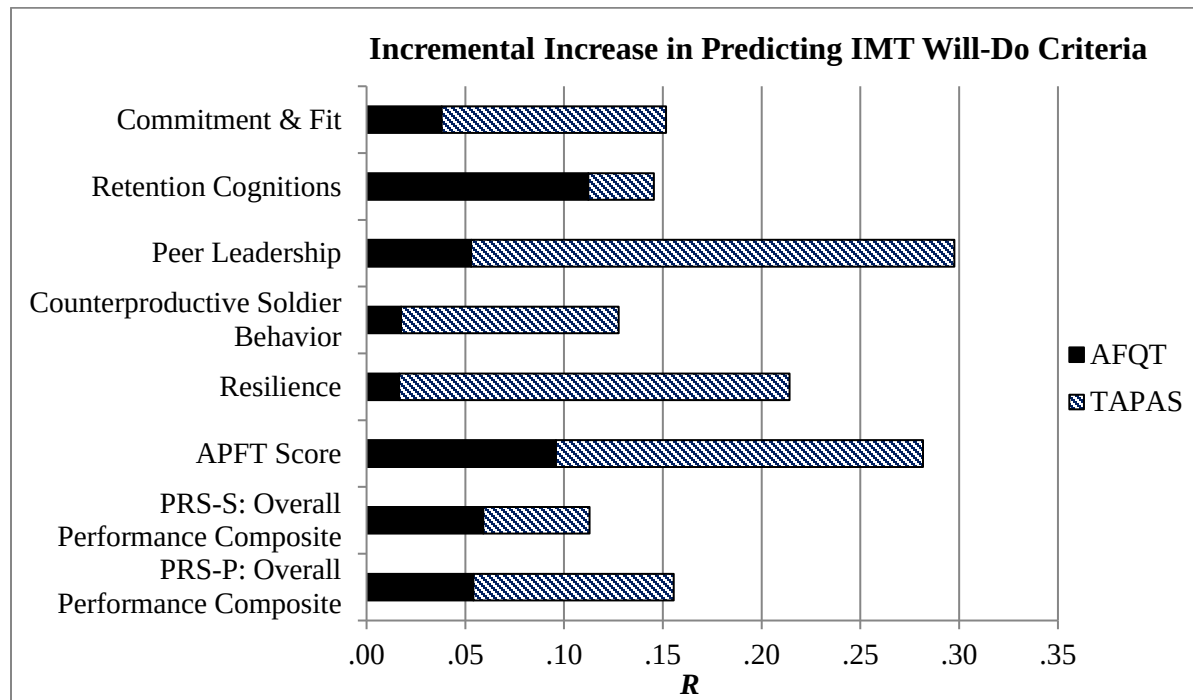


Figure 7.1. Increase in prediction of IMT criteria using Will-Do TAPAS scales.

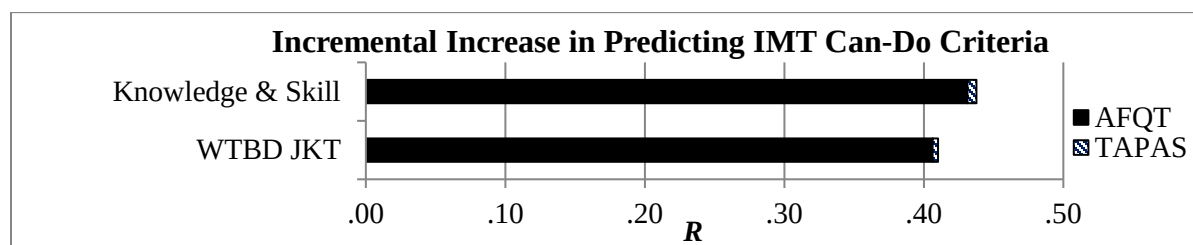


Figure 7.2. Increase in prediction of IMT criteria using Can-Do TAPAS scales.

Predicting In-Unit Performance

The incremental validity results for predicting in-unit performance are presented in Tables 7.13 to 7.15. 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 the AFQT alone. Results for which predictive gains (i.e., ΔR s) were greater than .05 are mentioned below.

For motivation-based criteria, the Will-Do composite showed increases beyond the AFQT in predicting the APFT Score ($\Delta R = .20$), Peer Leadership ($\Delta R = .16$), Resilience ($\Delta R = .14$), and PRS-P: Overall Performance ($\Delta R = .08$). With respect to predicting the Can Do outcomes of Knowledge and Skill or WTBD JKT scores, none of the TAPAS composite predictors demonstrated meaningful incremental validity beyond the AFQT (ΔR s $\leq .01$). Although the Can-Do and Adaptation composites showed statistically significant increases beyond the AFQT, these effects are very near zero.

The TAPAS Adaptation composite also demonstrated incremental validity for predicting APFT Score ($\Delta R = .13$) and Resilience ($\Delta R = .09$). Similar to the analyses of the IMT criteria, the Can-Do composite did not provide incremental validity in the prediction of any in-unit criteria. However, this result is expected given that the Can-Do composite is not intended to be related to Will Do outcomes, and the AFQT is an established strong predictor of Can Do outcomes.

The results of the incremental validity analyses for criterion-specific TAPAS scales are shown in Figures 7.3 and 7.4 for Will Do and Can Do criteria, respectively. Similar to the results of the TAPAS composite models predicting both IMT and in-unit criteria, the Will-Do scales provided the largest gains over the AFQT. Increases in R were largest for the APFT Score, Peer Leadership, and Resilience (ΔR s $> .17$). Results of the model including the Can-Do TAPAS scales exhibited negligible increments beyond AFQT in predicting either Knowledge and Skill or the WTBD JKT.

Table 7.13. Incremental Validity Estimates for the TAPAS over AFQT for Predicting In-Unit Technical Performance Criteria

In-Unit Criterion Measure / Model	AFQT R	AFQT + TAPAS R	ΔR
<i>Knowledge & Skill</i>			
		<i>n = 4,523 - 6,684</i>	
Can-Do ^a	.44	.45	.01
Will-Do	.44	.44	.00
Adaptation ^a	.44	.44	.00
<i>WTBD JKT</i>			
		<i>n = 4,517 - 6,677</i>	
Can-Do ^a	.43	.44	.01
Will-Do	.44	.44	.00
Adaptation ^a	.44	.44	.00

Note. WTBD JKT = Warrior Tasks and Battle Drills Job Knowledge Test. R = multiple correlations between the AFQT and selected TAPAS composite scales with the targeted criterion measure. ΔR = Increment in R from adding the selected TAPAS composite scale to the regression model [(AFQT + TAPAS) – AFQT Only]. Bolded values indicate $p < .05$.

^a Because the Can-Do and Adaptation composites are based on a subset of the data, these results may underestimate their incremental validity.

Table 7.14. Incremental Validity Estimates for the TAPAS over AFQT for Predicting In-Unit ALQ and APFT Criteria

In-Unit Criterion Measure / Model	AFQT R	AFQT + TAPAS R	ΔR
<i>Commitment & Fit</i> <i>n</i> = 5,422 - 7,734			
Can-Do ^a	.04	.05	.01
Will-Do	.03	.07	.04
Adaptation ^a	.03	.05	.01
<i>Retention Cognitions</i> <i>n</i> = 764 - 877			
Can-Do ^a	.14	.15	.01
Will-Do	.11	.12	.00
Adaptation ^a	.11	.14	.03
<i>Peer Leadership</i> <i>n</i> = 765 - 878			
Can-Do ^a	.04	.04	.01
Will-Do	.05	.21	.16
Adaptation ^a	.05	.06	.01
<i>Counterproductive Soldier Behavior</i> <i>n</i> = 765 - 878			
Can-Do ^a	.08	.08	.00
Will-Do	.10	.11	.00
Adaptation ^a	.10	.11	.01
<i>Resilience</i> <i>n</i> = 765 - 878			
Can-Do ^a	.02	.04	.01
Will-Do	.02	.16	.14
Adaptation ^a	.02	.11	.09
<i>APFT Score</i> <i>n</i> = 5,236 - 7,516			
Can-Do ^a	.03	.03	.00
Will-Do	.02	.22	.20
Adaptation ^a	.02	.15	.13

Note. R = multiple correlations between the AFQT and selected TAPAS composite scales with the targeted criterion measure. ΔR = Increment in R from adding the selected TAPAS composite scale to the regression model [(AFQT + TAPAS) – AFQT Only]. Bolded values indicate $p < .05$.

^a Because the Can-Do and Adaptation composites are based on a subset of the data, these results may underestimate their incremental validity.

Table 7.15. Incremental Validity Estimates for the TAPAS over AFQT for Predicting In-Unit Performance Rating Criteria

In-Unit Criterion Measure / Model	AFQT <i>R</i>	AFQT + TAPAS <i>R</i>	ΔR
<i>PRS-S: Overall Performance</i>		<i>n = 4,161 - 6,016</i>	
Can-Do ^a	.08	.08	.00
Will-Do	.08	.13	.04
Adaptation ^a	.08	.10	.01
<i>PRS-P: Overall Performance</i>		<i>n = 615 - 692</i>	
Can-Do ^a	.01	.05	.04
Will-Do	.01	.09	.08
Adaptation ^a	.01	.01	.00

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

^a Because the Can-Do and Adaptation composites are based on a subset of the data, these results may underestimate their incremental validity.

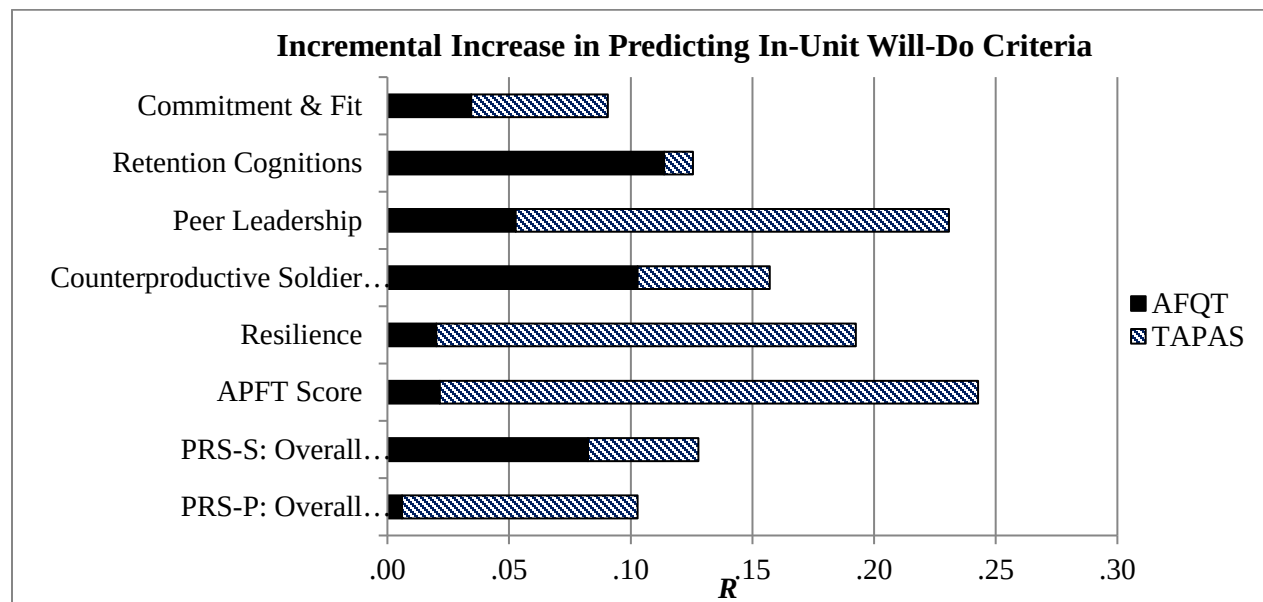


Figure 7.3. Increase in prediction of in-unit criteria using Will-Do TAPAS scales.

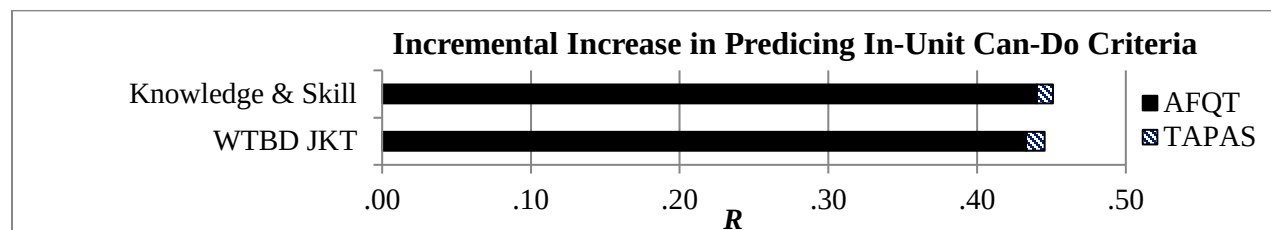


Figure 7.4. Increase in prediction of in-unit criteria using Can-Do TAPAS scales.

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 (both IMT and In-Unit), IMT Restarts, and attrition at 6, 12, and 24 months. 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) from the AFQT-only to the AFQT + TAPAS composite models.

Odds ratios can be used to assess the likelihood (or odds) of a given outcome depending on change in a predictor. Specifically, for a given logistic regression model, a unique odds ratio is estimated for each predictor, and represents the amount of change in the odds 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. Note that ORs equal to 1.0 reflect no relationship between a given predictor and outcome, ORs greater than 1.0 reflect positive relationships, and ORs between 0.0 and less than 1.0 reflect negative relationships (i.e., decreasing odds of the outcome with increasing values of the predictor). For ORs below 1.0, values closer to 0.0 indicate stronger negative relationships. 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 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 significantly better prediction of the outcome than the AFQT alone.

Results of the analyses examining Disciplinary Incidents (IMT and in-unit) and IMT Restarts are provided in Table 7.16. 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 ($OR_{Will-Do} = .988$; $OR_{Adaptation} = .993$) and IMT Restarts ($OR_{Will-Do} = .994$; $OR_{Adaptation} = .995$), 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 had significant relationships ($OR_{Will-Do} = .993$; $OR_{Adaptation} = .996$) and resulted in better model fit over the AFQT alone. The Can-Do composite did not predict either IMT or in-unit Disciplinary Incidents (as evidenced by CIs that include 1.000 for the associated ORs).

Table 7.17 presents the results of the logistic regression analyses examining attrition through 6, 12, and 24 months of service for Regular Army Soldiers. The Will-Do ($.990 \leq OR_{\text{Will-Do}} \leq .993$) and Adaptation ($.991 \leq OR_{\text{Adaptation}} \leq .992$) composites were negatively related to attrition at all three time points, and their respective inclusion in the models resulted in significantly better fit over the AFQT alone. Conversely, the TAPAS Can-Do composite was positively related to attrition at each of the time points for Regular Army Soldiers ($OR_{\text{Can-Do}} > 1.00$). However, the OR lower bounds of the CIs for the Can-Do composite were very near 1.0 in each of the attrition models, and these results may not represent a true effect.

Table 7.16. Incremental Validity Estimates for the TAPAS Composites over AFQT for Predicting Dichotomous Criteria

Criterion Measure / Model	OR_{AFQT} (CI)	OR_{TAPAS} (CI)	r_{pb}	$\Delta\text{-2LL}$
<i>IMT Disciplinary Incidents</i>				
<i>n = 49,620 - 58,652</i>				
Can-Do ^a				
AFQT	.964 (.944-.984)		.02	
AFQT+TAPAS	.973 (.952-.995)	.999 (.998-1.000)	.02	5.71
Will-Do				
AFQT	.963 (.945-.981)		.02	
AFQT + TAPAS	.982 (.963-1.001)	.988 (.987-.989)	.10	632.63
Adaptation				
AFQT	.963 (.945-.981)		.02	
AFQT + TAPAS	.987 (.968-1.006)	.993 (.992-.994)	.06	213.76
<i>IMT Restarts</i>				
<i>n = 465,945 - 593,580</i>				
Can-Do ^a				
AFQT	.974 (.962-.987)		.01	
AFQT+TAPAS	.980 (.966-.993)	.999 (.999-1.000)	.01	4.41
Will-Do				
AFQT	.978 (.967-.989)		.00	
AFQT + TAPAS	.987 (.976-.998)	.994 (.994-.995)	.03	401.84
Adaptation				
AFQT	.978 (.967-.989)		.00	
AFQT + TAPAS	.994 (.982-1.005)	.995 (.994-.995)	.02	303.29

Table 7.16. (Continued)

Criterion Measure / Model	OR _{AFQT} (CI)	OR _{TAPAS} (CI)	r_{pb}	Δ -2LL
<i>In-Unit Disciplinary Incidents</i> <i>n = 5,499 - 7,857</i>				
Can-Do ^a				
AFQT	.940 (.883-1.001)		.03	
AFQT+TAPAS	.947 (.884-1.014)	.999 (.996-1.002)	.03	0.27
Will-Do				
AFQT	.933 (.885-.983)		.03	
AFQT + TAPAS	.934 (.886-.985)	.993 (.990-.995)	.07	30.29
Adaptation				
AFQT	.933 (.885-.983)		.03	
AFQT + TAPAS	.943 (.894-.995)	.996 (.993-.999)	.04	9.41

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 values indicate significant change in model fit based on a Likelihood Ratio χ^2 test, $p < .05$.

^a Because the Can-Do composite is based on a subset of the data, these results may underestimate its incremental validity.

Table 7.17. Incremental Validity Estimates for the TAPAS Composite Scores over AFQT for Predicting Cumulative Attrition through 24 Months of Service (Regular Army Only)

Attrition Measure / Model	OR _{AFQT} (CI)	OR _{TAPAS} (CI)	r_{pb}	Δ -2LL
<i>6 Month</i> <i>n = 253,804 - 320,593</i>				
Can-Do ^a				
AFQT	.824 (.814-.835)		.06	
AFQT+TAPAS	.815 (.804-.827)	1.001 (1.001-1.002)	.06	17.59
Will-Do				
AFQT	.826 (.817-.836)		.06	
AFQT + TAPAS	.836 (.827-.846)	.990 (.989-.991)	.08	1252.36
Adaptation				
AFQT	.826 (.817-.836)		.06	
AFQT + TAPAS	.850 (.841-.860)	.991 (.990-.991)	.08	1039.72

Table 7.17. (Continued)

Attrition Measure / Model	OR _{AFQT} (CI)	OR _{TAPAS} (CI)	r_{pb}	Δ -2LL
12 Month				
<i>n</i> = 219,415 - 285,695				
Can-Do ^a				
AFQT	.841 (.831-.851)		.06	
AFQT+TAPAS	.831 (.820-.843)	1.001 (1.001-1.002)	.06	20.33
Will-Do				
AFQT	.843 (.834-.853)		.06	
AFQT + TAPAS	.853 (.844-.863)	.990 (.990-.991)	.09	1297.87
Adaptation				
AFQT	.843 (.834-.853)		.06	
AFQT + TAPAS	.868 (.859-.878)	.991 (.990-.991)	.08	1188.08
24 Month				
<i>n</i> = 164,618 - 229,487				
Can-Do ^a				
AFQT	.805 (.795-.815)		.08	
AFQT+TAPAS	.795 (.785-.806)	1.001 (1.001-1.002)	.08	20.95
Will-Do				
AFQT	.809 (.800-.817)		.08	
AFQT + TAPAS	.816 (.808-.825)	.993 (.992-.993)	.10	816.76
Adaptation				
AFQT	.809 (.800-.817)		.08	
AFQT + TAPAS	.828 (.819-.837)	.992 (.992-.993)	.10	893.11

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 values indicate significant change in model fit based on a Likelihood Ratio χ^2 test, $p < .05$.

^a Because the Can-Do composite is based on a subset of the data, these results may underestimate its incremental validity.

Figures 7.5 and 7.6 display the accuracy of the logistic regression models in predicting the dichotomous outcomes (Disciplinary Incidents, IMT Restarts, and attrition). 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 *c* statistic, or 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).¹³ The area under the curve (AUC) can range from .50 to 1.0, corresponding to 50% (or chance) and 100% accuracy, respectively. For example, the closer AUC is to 1.0, the more likely it is that Soldiers with lower TAPAS scores will have at least one IMT Disciplinary Incident.

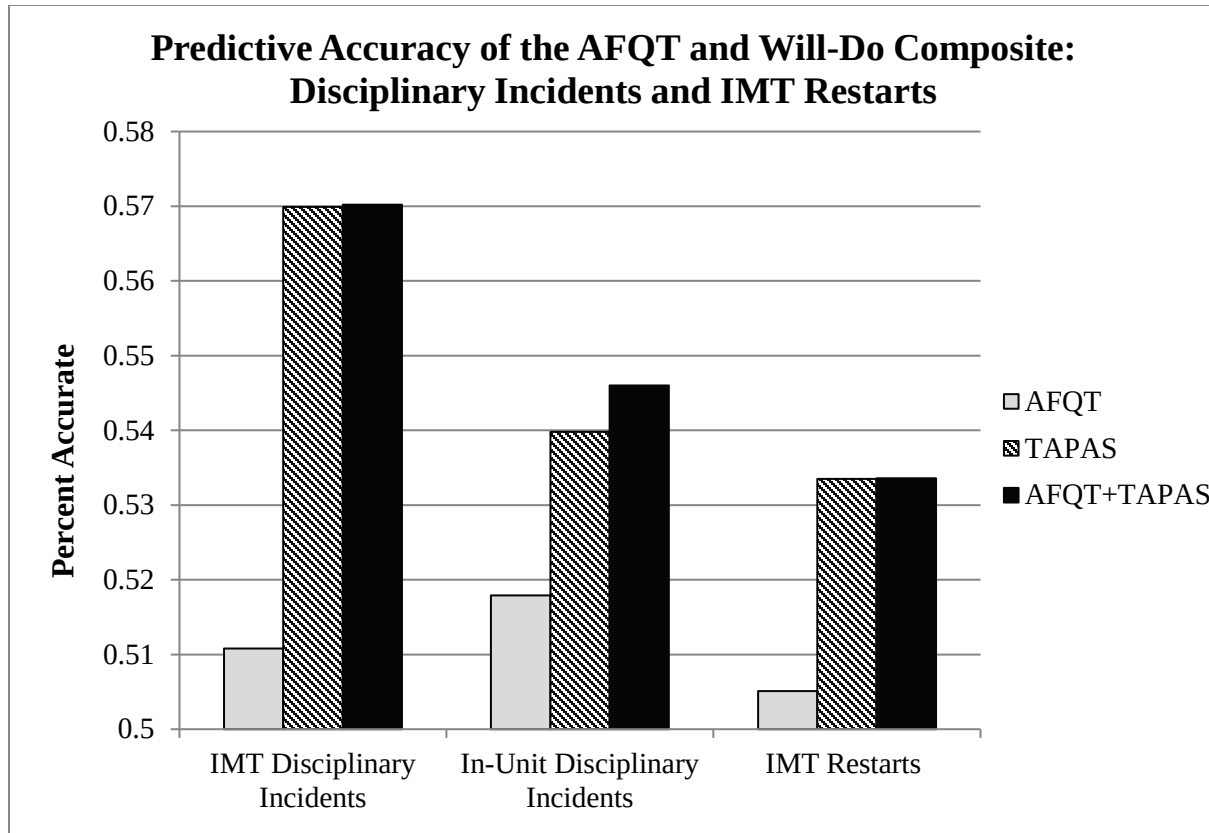


Figure 7.5. Predictive accuracy of the AFQT and Will-Do TAPAS composite in the discrimination of both IMT and In-Unit Disciplinary Incidents and IMT Restarts.

In particular, Figure 7.5 displays the results of the AFQT and Will-Do TAPAS composite models in predicting Disciplinary Incidents (IMT and in-unit) and IMT Restarts. For all three criteria, the probability of discriminating between Soldiers with and without incidence of the outcome of interest is lowest (less than 51.8%) when using the AFQT alone. However, the predictive accuracy increases for the Will-Do composite and AFQT + Will-Do composite models for all outcomes. Note the combined model demonstrated the most accuracy, ranging from 57.0% for IMT Disciplinary Incidents to 53.4% for IMT Restarts. In addition, the increase in discrimination accuracy for the combined model compared to the Will-Do composite-only model was generally small, suggesting that the AFQT adds little value to the discrimination of these dichotomous outcomes.

¹³ The ROC curve is a plot of sensitivity (i.e., probability of detecting a true positive) versus specificity (i.e., probability of detecting a true negative) across a range of potential cut scores for continuous predictors in a logistic regression model (Cook, 2007). For the purposes of evaluating the logistic regression models discussed in this report, only the area under the ROC curve (i.e., AUC) is presented and discussed.

Figure 7.6 displays the results of the AFQT and Adaptation TAPAS composite models in predicting attrition for Regular Army Soldiers. For 6-, 12-, and 24-month attrition, the combined AFQT + Adaptation composite model resulted in the highest discrimination accuracy (approximately 57% for each time point). When comparing the AFQT-only and Adaptation composite-only models, the Adaptation composite-only model evidenced higher accuracy than the AFQT-only at both 6- and 12-months. For each time point, the accuracy of the combined model improved by at least 1% over either predictor alone, suggesting that both the AFQT and Adaptation TAPAS composite contribute to the prediction of attrition.

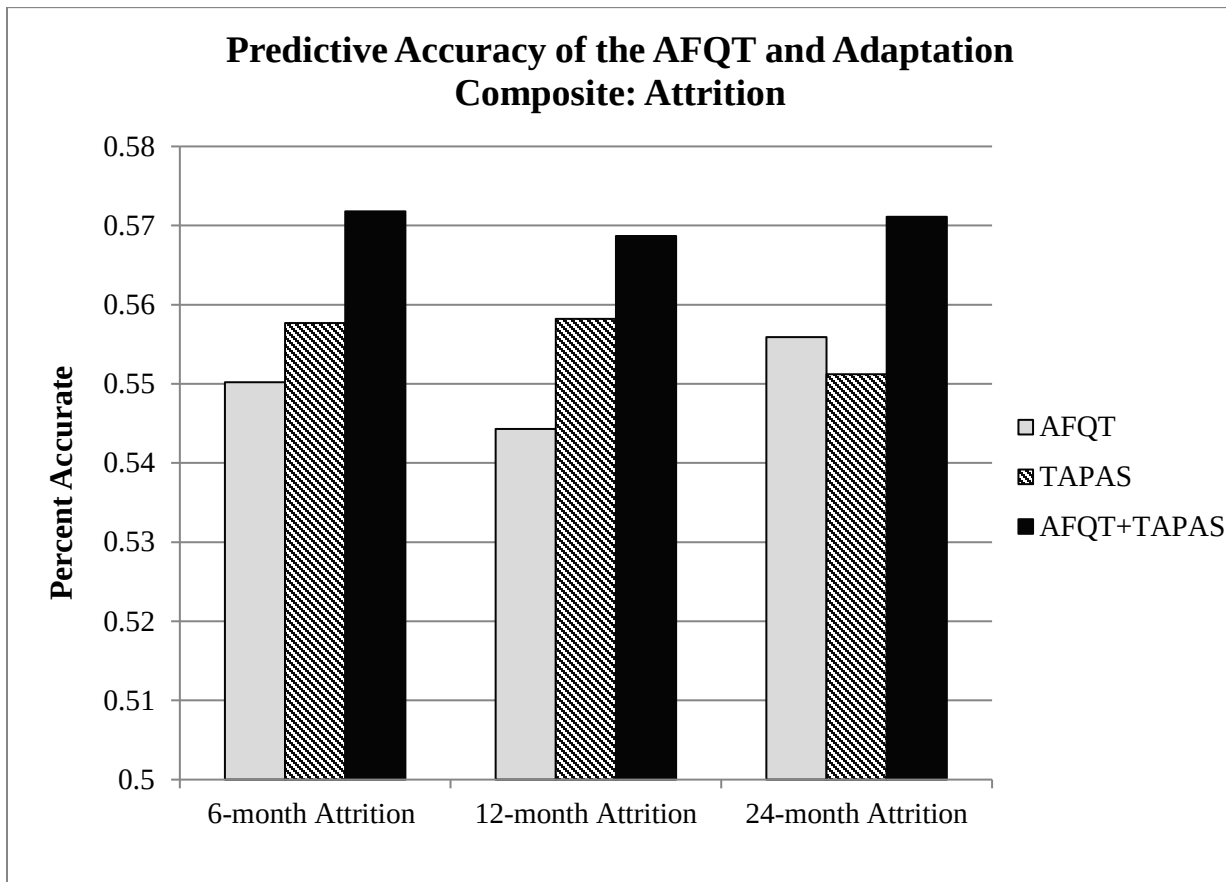


Figure 7.6. Predictive accuracy of the AFQT and Adaptation TAPAS composite in the discrimination of attrition outcomes for Regular Army Soldiers.

Summary

The TAPAS exhibited positive relationships with numerous criteria in both the meta- and incremental validity analyses. For both IMT and in-unit outcomes, the Will-Do TAPAS composite demonstrated the strongest relationships with APFT score, Peer Leadership, and Resilience. Results of the incremental validity analyses in particular showed that the Will-Do composite contributed beyond the AFQT in predicting these outcomes. Although not as strong, it also showed positive relationships with the performance rating scale Overall Performance composites for both the supervisors/cadre (PRS-S) and peer (PRS-P) measures. Regarding the

dichotomous outcomes, the Will-Do composite helped predict Disciplinary Incidents and IMT Restarts better than the AFQT alone. The Adaptation Composite also contributed in the prediction of attrition beyond the AFQT alone. Given the well-established strength of the AFQT to predict cognitive outcomes, it is unsurprising that the TAPAS does not help in the prediction of the knowledge-based outcomes. Nonetheless, the meta-analytic results showed that the Can-Do TAPAS composite was positively related to those criteria.

CHAPTER 8: SUMMARY AND A LOOK AHEAD

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

Summary of the TOPS IOT&E Method

To expand the basis on which applicants are evaluated for enlistment, the Army conducted an initial operational test and evaluation (IOT&E) of the *Tier One Performance Screen (TOPS)*. The assessment examined in TOPS, the Tailored Adaptive Personality Assessment System (TAPAS), was administered to most applicants testing at Military Entrance Processing locations.

To evaluate the TAPAS, we collected criterion data on Soldiers at multiple points during their time in service. For many Soldiers we were able to retrieve outcome data from administrative records, including training course grades, training completion rates, and separation status. Data from additional measures were collected from Soldiers in selected MOS as they completed initial military training (IMT). These measures included job knowledge tests (JKTs), an attitudinal assessment (the Army Life Questionnaire; ALQ), and performance rating scales (PRS). Versions of the JKTs and the ALQ suited for Soldiers in their first enlistment term were also administered to Soldiers after they joined their units. For most of the roughly 8-year IOT&E period, performance ratings were collected from Soldiers' training cadre (in IMT) and supervisors (in-units). Toward the end of the period, we shifted to collecting ratings from peers only at IMT and from both supervisors and peers in units. At 6-month intervals throughout the IOT&E period, we constructed analysis datasets incorporating TAPAS and criterion data and conducted evaluation analyses.

Although IOT&E data were collected through the end of CY2018, the data presented in this report are based on TAPAS and criterion data collected through the spring of 2018. In addition to reporting the most recent evaluation results, this report also stands as a capstone for the research conducted as part of the TOPS project since its inception in 2009. The foundational database consists of over a million (1,132,118) applicants who took the TAPAS; 1,048,245 of these individuals were in the TOPS Applicant Sample. The Applicant Sample (used for analysis purposes) excludes prior service applicants and those individuals ineligible for service based on education requirements or extremely low AFQT scores. The validation sample sizes are considerably smaller, with the IMT Validation Sample comprising 69,525 Soldiers, the In-Unit Validation Sample comprising 8,513 Soldiers, and the Administrative Validation Sample (which includes Soldiers with criterion data [e.g., attrition] from at least one administrative source) comprising 621,590 Soldiers.

Data from the JKTs, PRS, ALQ, and administrative sources were combined to yield an array of scores representing important Soldier outcomes. In general, the criterion scores exhibit acceptable psychometric properties and a sensible pattern of intercorrelations. The exception to this is the Army-wide and MOS-specific PRS, which have continued to exhibit low interrater reliability. Introduction of peer ratings improved reliability, but the psychometric characteristics of the ratings measures were still weaker than desired.

Summary of Evaluation Results

Evaluation results suggest that the TAPAS holds promise for new Soldier selection. As reported in Chapter 7, the TAPAS composites demonstrated incremental validity over the AFQT in predicting important first-term Soldier outcomes. In particular, the TAPAS Will-Do composite demonstrated the greatest incremental validity overall. Of the eight Will Do criteria, ΔR estimates for six IMT and three in-unit criteria exceeded .05. The TAPAS Adaptation composite also yielded incremental validity estimates above .05 for three IMT criteria and two in-unit criteria. In addition, both the Will-Do and Adaptation composites demonstrated negative relationships with the 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. Findings from the meta-analyses provided additional support for these observed relationships among the TAPAS composites and various outcomes. In general, the strongest meta-analytic corrected correlations were seen for outcomes where TAPAS provided the greatest prediction beyond the AFQT in the incremental validity analyses. Conversely, the TAPAS Can-Do composite did not provide any notable incremental validity for the criteria examined here. This finding is not surprising given the established strength of the AFQT in the prediction of cognitively-based criteria. Nonetheless, the Can-Do composite had small positive meta-analytic correlations with the cognitive outcomes.

As demonstrated in the last evaluation report (Knapp & Kirkendall, 2018), a distinction was also seen when comparing AFQT Category IIIB and IV Soldiers who scored above and below the 10th percentile on TAPAS. One particularly large difference was for disciplinary incidents where the low scoring groups had approximately 5% and 8% more disciplinary incidents, respectively, compared to the higher scoring IIIB and IV groups. The higher scoring Category IIIB and IV groups had lower attrition rates (ranging from a difference of 1.2% to 5.6%) than the low scoring groups at 6, 12, and 24 months' time in service. The Adaptation composite generally provided small incremental validity gains for predicting attrition, showing relatively larger gains for predicting attrition later in the enlistment term.

These findings are consistent with past evaluations in this series (Bynum & Mullins, 2017; Knapp & Heffner, 2011, 2012; Knapp et al., 2011; Knapp & Kirkendall, 2018; Knapp & LaPort, 2013a, 2013b, 2014) and the original research that led Army policy-makers to select TAPAS for the TOPS IOT&E (Knapp & Heffner, 2010).

Looking Ahead

Changes to Predictor Measures

In September of 2013, a third series of adaptive forms of the TAPAS was introduced at the MEPS. Each form measures 13 dimensions: the same 10 core dimensions plus three of eight experimental dimensions. The eight experimental dimensions assessed vary by version. In total, the newer versions of the TAPAS collectively measure 18 dimensions. The experimental dimensions will be evaluated for potential use in revised or new TAPAS composites, once sufficient data are available.

Changes to Criterion Measures

The IMT performance rating scales have been problematic for some time, in part because they have exhibited low reliability, but also because it is burdensome for IMT cadre to complete the ratings. In addition, IMT classes are often so large that the cadre completing the PRS are overwhelmed with the number of Soldiers they are asked to rate and are rarely familiar with all of their ratees. Collecting performance ratings from peers is a strategy that could provide a different perspective on Soldier performance and potentially more reliable data since it would be more feasible to gather data from multiple raters per Soldier. Therefore, ARI has developed peer versions of the Army-wide PRS that were introduced into criterion data collections in both IMT and in-unit starting in December 2016. To reduce burden on operations, collection of cadre ratings was suspended in May 2017 but supervisor ratings of Soldiers in-unit continued throughout the TOPS IOT&E.

The IMT and in-unit versions of the ALQ have been revised to reduce redundancy and make room for addition of new outcomes of interest to the Army. New topics cover content related to organizational citizenship behaviors, counterproductive work behaviors, work-life balance, deployment satisfaction, reasons for choosing to join the Army/specific MOS, and resilience. After pilot testing with Soldiers outside the TOPS sample, the new ALQs were introduced into the criterion data collections in May 2017. As the assessments used for selection and assignment in the military change, the ALQ will continue to evolve.

Related Research

Though the TOPS IOT&E is ending, a follow on project entitled Validation of Accessions Screening Tools (VAST) has already begun. VAST will pick up where TOPS left off and continue to validate the potential of TAPAS for accessions testing. This follow on project will also continue to report on the psychometric properties of the PRS and ALQ as additional data are collected. This updated criterion and validation information will be published in a series of VAST technical reports and research notes.

ARI is conducting research focused on evaluating and improving the psychometric characteristics of the TAPAS and facilitating its potential use for operational administration. Future research on this assessment will focus on further refining the content and developing new TAPAS dimensions that can be used to predict a broader range of outcomes in the U.S. Military.

ARI also has several lines of research regarding MOS assignment. Currently the Army primarily relies on aptitude area composites derived from ASVAB subtest scores for MOS assignment, but measures of temperament and interests have also shown to be predictive of attrition and job attitudes for a subset of MOS included in the TOPS research (Ingerick, Diaz, & Putka, 2009; Knapp, Owens, & Allen, 2012; Nye et al., 2012). In light of these promising results, ARI is developing and evaluating the utility of a new interest assessment (the Adaptive Vocational Interest Diagnostic [AVID]) for job assignment. ARI's goal is to use a holistic approach to selection and assignment, so future research will help determine the best way to combine ASVAB, TAPAS, AVID, and possibly additional predictor measures to optimize the MOS assignment process.

Conclusion

Inclusion of non-cognitive measures in initial entry screening allows the Army to predict a broader range of valued Army outcomes than traditional cognitive ability measures and educational credential screening. The TAPAS, specifically, demonstrates the ability to predict an expanded concept of Soldier performance beyond AFQT to include motivation, disciplinary behavior, adaptability, adjustment to military life, and to some extent, attrition. Given the importance of improving prediction of outcomes not associated with cognitive ability, additional research should continue to refine and expand the prediction potential of TAPAS and other experimental non-cognitive pre-enlistment measures.

REFERENCES

- Allen, M. T., Cheng, Y. A., Putka, D. J., Hunter, A., & White, L. (2010). Analysis and findings. In D.J. Knapp & T.S. Heffner (Eds.). *Expanded enlistment eligibility metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267) (pp 29-51). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Allen, M. T., Knapp, D. J., & Owens, K. S. (2013). *Validating future force measures (Army Class): Concluding analyses* (Technical Report in preparation). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Amit, K., Lisak, A., Popper, M., & Gal, R. (2007), Motivation to lead: research on the motives for undertaking leadership roles in the Israel defense forces (IDF). *Military Psychology*, 19, 137-160.
- Bateman, T. S., & Organ, D. W. (1983). Job satisfaction and the good soldier: The relationship between affect and employee "citizenship." *Academy of Management Journal*, 26, 587-595.
- Bonanno, G. (2004). Loss, trauma, and human resilience: Have we underestimated the human capacity to thrive after extremely aversive events? *American Psychologist*, 59, 20-28.
- Bynum, B. H., & Beatty, A. S. (2014). Description and psychometric properties of criterion measures. In D. J. Knapp & K. A. LaPort (Eds.), *Tier One Performance Screen initial operational test and evaluation: 2012 annual report* (Technical Report 1342) (pp. 15-31). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Bynum, B. H., & Mullins, H. M. (Eds.) (2017). *Tier One Performance Screen Initial operational test and evaluation: 2013 annual report* (Technical Report 1359). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- 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). Hillsdale, NJ: Erlbaum.
- Campbell, J. P., & Knapp, D. J. (Eds.) (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: 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 & Associates (Eds.), *Personnel Selection in Organizations*. San Francisco CA: 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.
- Chan, K. Y., & Drasgow, F. (2001). Toward a theory of individual differences and leadership: Understanding the motivation to lead. *Journal of Applied Psychology*, 86, 481-498.

- Chernyshenko, O. S., & Stark, S. (October, 2007). *Criterion validity evidence for narrow temperament clusters: A meta-analysis of military studies*. Paper presented at the 49th annual conference of the International Military Testing Association. Gold Coast, AU.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd edition)*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Collins, M., Le, H., & Schantz, L. (2005). Job knowledge criterion tests. 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. 49-58). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115, 928-935.
- Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology*, 90, 1241-1255.
- Dalal, R. S., Lam, H., Weiss, H. M., Welch, E. R., & Hulin, C. L. (2009). A within-person approach to work behavior and performance: Concurrent and lagged citizenship-counterproductivity associations, and dynamic relationships with affect and overall job performance. *Academy of Management Journal*, 52, 1051-1066.
- Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB) (FR-06-25)*. Alexandria, VA: Human Resources Research Organization.
- 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). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Drasgow, F., Whetzel, D. L., & Oppler, S. H. (2007). Strategies for test validation and refinement. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 349-384). New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.
- Ford, L. A. Campbell, R. C., Campbell, J. P., Knapp, D. J., & Walker, C. B. (2000). *21st century soldiers and noncommissioned officers: Critical predictors of performance* (Technical Report 1102). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Gruys, M. L., & Sackett, P. R. (2003). Investigating the dimensionality of counterproductive work behavior. *International Journal of Selection and Assessment*, 11, 30-41.

- Hoffman III, R. R., Muraca, S. T., Heffner, T. S., Hendricks, R., & Hunter, A. E. (2009). *Selection for accelerated basic combat training* (Technical Report 1241). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
- Ingerick, M., Diaz, T., & Putka, D. (2009). *Investigations into Army enlisted classification systems: Concurrent validation report* (Technical Report 1244). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Kemery, E. R., Dunlap, W. P., & Griffeth, R. W. (1988). Correction for variance restriction in point-biserial correlations. *Journal of Applied Psychology*, 73, 688-691.
- Kiger, T. B., Caramagno, J. P., & Reeder, M. C. (2017). Description and Psychometric Properties of Criterion Measures. In B. H. Bynum & H. M. Mullins (Eds.) *Tier One Performance Screen initial operational test and evaluation: 2013 annual report* (Technical Report 1359). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., Burnfield, J. L., Sager, C. E., Waugh, G. W., Campbell, J. P., Reeve, C. L., Campbell, R. C., White, L. A., & Heffner, T. S. (2002). *Development of predictor and criterion measures for the NCO21 Research program* (Technical Report 1128). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., & Heffner, T. S. (Eds.) (2009). *Predicting Future Force Performance (Army Class): End of Training Longitudinal Validation* (Technical Report 1257). Arlington, VA: 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). Arlington, VA: 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). Arlington, VA: 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). Fort Belvoir, VA: 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). Arlington, VA: 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*. Fort Belvoir, VA: 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). Fort Belvoir, VA: 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). Fort Belvoir, VA: 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). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D. J., Owens, K. S., & Allen, M. T. (Eds.) (2012). *Validating future force performance measures (Army Class): In-unit performance longitudinal validation* (Technical Report 1314). Arlington, VA: 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). Arlington, VA: 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). Arlington, VA: 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). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Moriarty, K. O., & Bynum, B. H. (2011). Description and psychometric properties of criterion measures. In D. J. Knapp & T. S. Heffner (Eds.), *Tier One Performance Screen initial operational test and evaluation: 2010 annual report* (Technical Report 1296) (pp. 20-28). Fort Belvoir: U.S. Army Research Institute for the Behavioral and Social Sciences.
- 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). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology*, 79, 475-480.

- 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*. Fort Belvoir, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.
- Peterson, N. G., Anderson, L. E., Crafts, J. L., Smith, D. A., Motowidlo, S. J., Rosse, R. L., Waugh, G. W., McCloy, R. A., Reynolds, D. H., & Dela Rosa, M. R. (1999). *Expanding the concept of quality in personnel: Final report* (Research Note 99-31). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Podsakoff, P. M., Ahearne, M., & MacKenzie, S. B. (1997). Organizational citizenship behavior and the quantity and quality of work group performance. *Journal of Applied Psychology*, 82, 262-270.
- Pulakos, E. D. (2007). Performance measurement. In D. L. Whetzel and G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 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.
- Putka, D. J., Hoffman, B. J., & Carter, N. T. (2014). Correcting the correction: When individual raters offer distinct but valid perspectives. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 7, 546-551.
- Putka, D. J., & Van Iddekinge, C. H. (2007). Work Preferences Survey. In D. J. Knapp & T. R. Tremble (Eds.), *Concurrent validation of experimental Army enlisted personnel selection and classification measures* (Technical Report 1205) (pp. 135-155). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Riegelhaupt, B. J., Harris, C. D., & Sadacca, R. (1987). *The development of administrative measures as indicators of soldier effectiveness* (Technical Report 754). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal*, 38, 555-572.
- Russell, T. L., & Sellman, W. S. (Eds.) (2009). *Development and pilot testing of an information and communications technology literacy test for military enlistees: Volume I technical report* (FR 08-128). Alexandria, VA: Human Resources Research Organization.

- Sparks, T. E., & Peddie, C. (2013). Description and psychometric properties of criterion measures. In D. J. Knapp & T. S. Heffner (Eds.), *Tier One Performance Screen initial operational test and evaluation: 2012 interim report* (Technical Report 1332) (pp. 15-31). Fort Belvoir: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Spector, P. E., Fox, S., Penney, L. M., Bruursema, K., Goh, A., & Kessler, S. (2006). The dimensionality of counterproductivity: Are all counterproductive behaviors created equal? *Journal of Vocational Behavior*, 68, 446 – 460.
- Stark, S. E., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise preference model. *Applied Psychological Measurement*, 29, 184-201.
- Stark, S. E., Chernyshenko, O. S., & Drasgow, F. (2010). Tailored adaptive personality assessment system (TAPAS-95s). In D. J. Knapp & T. S. Heffner (Eds.) *Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267) (pp 29-51). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Stark, S. E., Chernyshenko, O. S., & Drasgow, F. (2012). Adaptive testing with multi-unidimensional pairwise preference Items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*, 15, 463-487.
- Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology*, 26, 153-164.
- Strickland, W. J. (Ed.) (2005). *A longitudinal examination of first term attrition and reenlistment among FY1999 enlisted accessions* (Technical Report 1172). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.
- 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). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- 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). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

- Van Iddekinge, C. H., Roth, P. L., Putka, D. J., & Lanivich, S. E. (2011). Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. *Journal of Applied Psychology*, 96, 1167-1194.
- White, L. A., & Young, M. C. (1998, August). *Development and validation of the Assessment of Individual Motivation (AIM)*. Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.
- Williams, C. R. (1990). Deciding, when, how, and if to correct turnover correlations. *Journal of Applied Psychology*, 75, 732-737.
- Yukl, G., Gordon, A., & Taber, T. (2002). A hierarchical taxonomy of leadership behavior: integrating a half century of behavior research. *Journal of Leadership & Organizational Studies*, 9, 15-32.

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
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	Inter-rater 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
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 the TAPAS Scales and Composites

TAPAS Composite/ TAPAS Scale	15D-Static/CAT v4/5/7/8 (<i>n</i> = 100,618-421,900) ^a		13D-CAT v9 (<i>n</i> = 197,179)		13D-CAT v10 (<i>n</i> = 197,017)		13D-CAT v11 (<i>n</i> = 198,710)	
	<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>Individual TAPAS Scales</i>								
Achievement	-0.01	0.98	-.01	.98	-.04	.98	-.01	.99
Adjustment	-0.02	0.98	—	—	—	—	—	—
Adventure Seeking	-0.06	0.98	—	—	—	—	—	—
Attention Seeking	-0.04	0.98	—	—	—	—	-.01	.97
Commitment to Serve	0.00	0.98	-.04	.98	—	—	-.03	1.00
Cooperation	0.01	0.98	-.02	.99	—	—	—	—
Courage	-0.02	0.98	—	—	.00	.99	—	—
Dominance	-0.02	0.98	-.02	.97	-.06	.97	-.03	.99
Even Tempered	0.01	0.98	.02	1.01	-.01	1.01	.04	1.01
Intellectual Efficiency	-0.02	0.97	.03	.97	.00	.95	.04	.97
Non-Delinquency	0.03	0.98	—	—	.01	.98	—	—
Optimism	0.00	0.98	.00	.99	.01	.98	.00	.99
Order	0.01	0.98	-.02	.96	-.03	.97	-.01	.97
Physical Conditioning	0.01	0.98	-.04	1.01	-.03	.98	-.01	1.01
Responsibility	0.00	0.98	.00	.99	—	—	—	—
Self-Control	0.00	0.98	—	—	—	—	—	—
Selflessness	0.01	0.98	.06	1.00	.07	.99	.03	.99
Situational Awareness	-0.02	0.98	—	—	-.08	.99	—	—
Sociability	-0.01	0.98	.03	1.01	.03	1.02	.02	1.00
Team Orientation	-0.01	0.99	—	—	—	—	.03	1.00
Tolerance	0.00	0.98	.09	.99	.09	.99	.11	.98
<i>TAPAS Composites</i>								
Can-Do	99.79	20.93	100.05	19.68	99.42	19.67	100.18	19.63
Will-Do	99.93	19.16	99.33	20.22	98.97	19.61	99.58	20.32
Adaptation	100.19	19.57	99.58	19.96	99.47	19.84	100.09	19.84

Note. Results are limited to the TOPS Applicant Sample (non-prior service, Education Tier 1 and 2, and AFQT ≥ 10) with valid TAPAS score data on forms other than the original 13D-CAT which was excluded from the Applicant Sample.

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

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

TAPAS Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. AFQT																					
<i>Individual TAPAS Scales</i>																					
2. Achievement	.05																				
3. Adjustment	.12	.10																			
4. Adventure Seeking	.10	.10	.15																		
5. Attention Seeking	.03	.05	.11	.17																	
6. Commitment to Serve	-.13	.22	.05	.04	.04																
7. Cooperation	-.08	.11	.07	-.14	-.03	.06															
8. Courage	.09	.27	.16	—	.10	—	—														
9. Dominance	.11	.33	.10	.13	.24	.14	-.06	.26													
10. Even Tempered	.09	.16	.22	-.05	-.04	.13	.31	.11	.00												
11. Intellectual Efficiency	.31	.30	.19	.07	.09	.09	.01	.23	.30	.13											
12. Non-Delinquency	-.04	.21	.01	-.17	-.14	.12	.23	.07	-.01	.25	.05										
13. Optimism	.09	.18	.25	.02	.09	.11	.15	.09	.13	.22	.13	.13									
14. Order	-.16	.21	-.07	-.09	-.05	.09	.10	.01	.09	.04	.10	.14	.03								
15. Physical Conditioning	.05	.22	.05	.25	.09	.10	-.05	.15	.19	-.05	.08	-.04	.06	.07							
16. Responsibility	.14	.37	.12	—	-.05	.15	.15	.15	.22	.22	.23	.23	.21	.17	.09						
17. Self-Control	-.02	.23	.09	—	-.10	—	.14	.09	.05	.22	.17	.26	.09	.19	-.04	.22					
18. Selflessness	-.06	.18	-.04	-.04	-.05	.09	.24	.10	.07	.15	.05	.18	.10	.10	.00	.23	.09				
19. Situational Awareness	.02	.22	.15	.10	.04	.07	.00	.20	.15	.15	.29	.13	.11	.18	.08	—	—	.05			
20. Sociability	-.11	.13	.10	—	.35	.15	.12	.12	.25	.08	.10	.01	.15	.02	.05	.09	-.07	.17	.05		
21. Team Orientation	-.09	.11	.05	—	.14	.12	—	.03	.13	.11	.00	.05	.10	.05	.07	.02	.05	.19	—	.25	
22. Tolerance	.06	.12	.03	—	.05	.07	.14	.09	.06	.17	.14	.07	.10	.04	-.04	.12	.11	.28	.10	.20	.12

Note. Results are limited to the Applicant Sample (Non-prior service, Education Tier 1 and 2, AFQT Category IV and above) with valid TAPAS score data, $n = 100,618 - 1,012,838$. Not all TAPAS scales were administered in every version; missing correlations indicate that the scales 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

Measure	Male		Female		<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
<i>AFQT</i>	56.00	22.63	49.57	21.48	-0.29
<i>Individual TAPAS Scales</i>					
Achievement	-0.04	0.99	0.06	0.96	0.10
Adjustment	0.04	0.97	-0.27	0.99	-0.32
Adventure Seeking	0.02	0.98	-0.28	0.96	-0.30
Attention Seeking	-0.01	0.98	-0.10	0.97	-0.10
Commitment to Serve	-0.03	1.00	-0.03	0.96	0.00
Cooperation	-0.04	0.98	0.14	0.98	0.18
Courage	0.04	0.98	-0.18	1.00	-0.22
Dominance	-0.03	0.99	-0.03	0.95	0.00
Even Tempered	0.01	0.99	0.01	1.00	0.00
Intellectual Efficiency	0.02	0.97	-0.05	0.94	-0.08
Non-Delinquency	-0.02	0.98	0.16	0.95	0.18
Optimism	-0.01	0.98	0.03	0.98	0.04
Order	-0.07	0.96	0.19	1.00	0.27
Physical Conditioning	0.07	0.99	-0.27	0.97	-0.34
Responsibility	-0.03	0.99	0.08	0.97	0.11
Self-Control	-0.01	0.98	0.05	0.98	0.06
Selflessness	-0.07	0.97	0.39	0.95	0.48
Situational Awareness	-0.02	0.98	-0.23	0.99	-0.21
Sociability	-0.01	1.00	0.07	1.00	0.09
Team Orientation	0.04	0.99	-0.09	0.99	-0.14
Tolerance	-0.01	0.98	0.31	0.95	0.33
<i>TAPAS Composites</i>					
Can-Do	101.11	20.06	95.81	19.38	-0.27
Will-Do	100.29	19.86	97.09	18.98	-0.16
Adaptation	101.85	19.40	93.59	19.56	-0.42

Note. Results are limited to the TOPS Applicant Sample (non-prior service, Education Tier 1 and 2, and AFQT $\geq$ 10) with valid TAPAS score data. Females are the referent group. Sample size for Females range from 21,404 to 224,332; for Males 75,416 to 767,464. 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

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>	60.34	21.87	43.89	19.23	-0.78	47.70	20.70	-0.59	59.86	23.80	-0.02
<i>Individual TAPAS Scales</i>											
Achievement	0.04	1.00	-0.04	0.95	-0.09	-0.08	0.97	-0.12	-0.27	0.97	-0.32
Adjustment	0.05	0.99	-0.12	0.95	-0.17	-0.12	0.96	-0.18	-0.24	0.96	-0.30
Adventure Seeking	0.15	0.97	-0.49	0.91	-0.67	-0.05	0.91	-0.21	-0.17	0.92	-0.33
Attention Seeking	-0.06	1.00	0.07	0.95	0.14	-0.03	0.95	0.03	-0.03	0.95	0.03
Commitment to Serve	0.02	0.99	-0.08	0.97	-0.09	-0.03	0.99	-0.05	-0.20	1.01	-0.22
Cooperation	-0.05	0.99	0.13	0.98	0.19	0.00	0.97	0.06	0.07	0.97	0.12
Courage	0.12	0.98	-0.19	0.99	-0.32	-0.04	0.97	-0.17	-0.37	1.00	-0.51
Dominance	0.00	1.02	0.01	0.90	0.01	-0.08	0.95	-0.08	-0.30	0.94	-0.29
Even Tempered	0.04	1.01	0.02	0.98	-0.02	-0.02	0.97	-0.06	-0.14	0.95	-0.17
Intellectual Efficiency	0.04	0.99	0.02	0.91	-0.01	-0.07	0.94	-0.11	-0.16	0.97	-0.20
Non-Delinquency	0.03	1.00	0.08	0.97	0.05	-0.01	0.95	-0.04	-0.11	0.93	-0.13
Optimism	0.03	0.99	0.00	0.98	-0.03	-0.02	0.97	-0.05	-0.16	0.96	-0.19
Order	-0.13	0.98	0.18	0.96	0.32	0.07	0.93	0.21	0.14	0.93	0.28
Physical Conditioning	0.05	1.02	-0.12	0.95	-0.18	-0.03	0.98	-0.08	-0.14	0.97	-0.19
Responsibility	0.09	0.98	-0.03	0.98	-0.13	-0.17	0.99	-0.27	-0.30	1.00	-0.41
Self-Control	-0.07	0.98	0.17	0.96	0.24	0.05	0.97	0.12	0.02	0.96	0.09
Selflessness	-0.03	1.00	0.20	0.97	0.22	0.02	0.97	0.05	0.09	0.94	0.12
Situational Awareness	-0.05	1.00	-0.05	0.97	0.00	-0.10	0.97	-0.05	-0.17	0.99	-0.12
Sociability	-0.01	1.02	0.08	0.97	0.09	0.01	0.98	0.02	-0.09	0.98	-0.08
Team Orientation	-0.02	1.00	0.03	1.00	0.05	0.06	0.99	0.08	0.18	1.01	0.20
Tolerance	-0.04	1.01	0.16	0.93	0.21	0.20	0.94	0.25	0.29	0.91	0.34
<i>TAPAS Composites</i>											
Can-Do	102.41	20.48	97.43	18.60	-0.25	96.80	19.20	-0.28	94.12	19.45	-0.41
Will-Do	101.09	20.25	98.11	18.44	-0.15	98.43	19.16	-0.13	93.60	19.26	-0.37
Adaptation	102.40	19.82	95.99	19.11	-0.33	98.52	19.35	-0.20	95.72	19.23	-0.34

Note. Results are limited to the TOPS Applicant Sample (non-prior service, Education Tier 1 and 2, and AFQT ≥ 10) with valid TAPAS score data. The Minority group is the referent group. Sample size for Whites range from 54,437 to 541,876; Blacks from 24,355 to 231,967; Hispanics from 15,542 to 165,468; Asians from 3,919 to 47,448. 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

CORRELATIONS AMONG CRITERION MEASURES IN THE IMT AND IN-UNIT VALIDATION SAMPLES

Table C.1. Correlations among the Army Life Questionnaire (ALQ) Scales in the IMT and In-Unit Validation Samples

Composite/Scale	1	2	3	4	5	6	7	8	9	10	11
1. Commitment and Fit		.41	.60	.84	.60	.86	.26	-.25	.61	.75	.64
2. Peer Leadership	.42		.26	.41	.27	.32	.05	-.06	.37	.25	.28
3. Retention Cognitions	.52	.23		.57	.97	.60	.19	-.29	.52	.28	.36
4. Affective Commitment	.87	.43	.53		.57	.66	.20	-.19	.57	.40	.47
5. Army Career Intentions	.49	.22	.97	.51		.61	.19	-.29	.52	.29	.37
6. Army Fit	.86	.28	.52	.72	.49		.27	-.34	.55	.44	.49
7. Army Life Adjustment	.54	.08	.33	.41	.31	.63		-.04		.18	.13
8. Counterproductive Soldier Behavior	-.33	-.03	-.17	-.20	-.17	-.38	-.36		-.19	-.08	-.10
9. Deployment Satisfaction										.40	.46
10. MOS Fit	.79	.29	.27	.47	.25	.49	.35	-.24			.61
11. MOS Satisfaction											
12. Motivation to Lead - Affective	.38	.81	.22	.37	.21	.29	.17	-.11		.26	
13. Motivation to Lead - Noncalculative	-.30	.22	-.16	-.14	-.14	-.39	-.46	.38		-.22	
14. Motivation to Lead - Social Normative	.58	.72	.28	.51	.26	.51	.28	-.28		.41	
15. Organizational Citizenship Behavior/Leadership	.42	.63	.24	.35	.23	.37	.29	-.14		.30	
16. Reenlistment Intentions	.50	.22	.96	.51	.86	.50	.33	-.16		.26	
17. Resilience	.54	.48	.26	.45	.25	.49	.32	-.29		.37	
18. Work/Life Balance	.41	.22	.32	.38	.32	.41	.19	-.09		.21	
19. Disciplinary Incidents (#)	-.09	-.06	-.04	-.06	-.04	-.08	-.15	.09		-.07	
20. Disciplinary Incidents (Y/N)	-.08	-.06	-.04	-.05	-.04	-.07	-.15	.08		-.06	
21. APFT Score	.07	.12	.03	.03	.02	.08	.22	-.05		.06	

Note. Correlations below the diagonal reflect the IMT ALQ, $n = 16,316$ -63,125. Correlations above the diagonal reflect the in-unit ALQ, $n = 978$ -8,284. Missing values reflect the scales that were not administered in either IMT or In-Unit. Correlations in bold are statistically significant, $p < .05$ (two-tailed).

Table C.1. (Continued)

Scale	12	13	14	15	16	17	18	19	20	21
1. Commitment and Fit	.35	-.16	.48	.38	.56	.47	.53	-.17	-.16	.08
2. Peer Leadership	.81	.20	.74	.70	.23	.47	.21	-.09	-.11	.19
3. Retention Cognitions	.24	-.16	.34	.24	.97	.29	.37	-.13	-.14	.11
4. Affective Commitment	.33	-.06	.44	.33	.53	.41	.43	-.12	-.13	.07
5. Army Career Intentions	.26	-.16	.35	.24	.88	.29	.36	-.13	-.13	.07
6. Army Fit	.29	-.25	.46	.35	.57	.46	.52	-.19	-.16	.09
7. Army Life Adjustment	.10	-.20	.11	.17	.21	.03	-.02	-.05	-.05	.14
8. Counterproductive Soldier Behavior	-.12	.33	-.30	-.11	-.26	-.28	-.19	.14	.15	-.09
9. Deployment Satisfaction	.34	-.17	.40	.41	.49	.52	.35	-.16	-.14	.14
10. MOS Fit	.22	-.08	.27	.24	.26	.26	.34	-.10	-.09	.04
11. MOS Satisfaction	.23	-.02	.27	.23	.35	.31	.59	-.11	-.10	.03
12. Motivation to Lead - Affective		-.13	.63	.50	.21	.40	.15	-.08	-.11	.21
13. Motivation to Lead - Noncalculative	-.10		-.30	-.25	-.15	-.18	-.03	.06	.06	-.09
14. Motivation to Lead - Social Normative	.61	-.28		.55	.31	.51	.26	-.14	-.16	.16
15. Organizational Citizenship Behavior/Leadership	.42	-.29	.49		.22	.48	.17	-.06	-.09	.19
16. Reenlistment Intentions	.21	-.15	.27	.24		.28	.36	-.11	-.12	.11
17. Resilience	.40	-.20	.56	.46	.25		.34	-.17	-.19	.20
18. Work/Life Balance	.16	-.01	.21	.19	.30	.33		-.11	-.13	.02
19. Disciplinary Incidents (#)	-.07	.05	-.06	-.07	-.04	-.08	-.06		.80	-.07
20. Disciplinary Incidents (Y/N)	-.07	.04	-.06	-.06	-.04	-.08	-.06	.86		-.09
21. APFT Score	.13	-.08	.11	.13	.03	.13	-.01	-.13	-.16	

Note. Correlations below the diagonal reflect the IMT ALQ, $n = 16,331-63,144$. Correlations above the diagonal reflect the in-unit ALQ, $n = 110-8,153$. Missing values reflect the scales that were not administered in either IMT or In-Unit. Correlations in bold are statistically significant, $p < .05$ (two-tailed).

Table C.2. Correlations among the Supervisor Performance Rating Scales (PRS-S) in the IMT Validation Sample

Domain/PRS	1	2	3	4	5	6	7
<i>Army-Wide</i>							
1. Overall Performance Composite							
2. Adjustment to the Army	.85						
3. Effort and Discipline	.85	.69					
4. MOS Qualification Knowledge	.84	.70	.64				
5. Physical Fitness & Bearing	.79	.64	.63	.58			
6. Working Effectively with Others	.85	.69	.72	.65	.60		
7. Overall Performance Rating	.81	.67	.64	.66	.58	.61	
<i>MOS-Specific</i>							
8. Infantryman (11B/C/X + 18X)	.78	.67	.60	.64	.58	.62	.54
9. Combat Engineer (12B)	.78	.59	.44	.74	.60	.55	.70
10. Fire Support Specialist (13F)	.82	.66	.59	.74	.57	.64	.73
11. Cavalry Scout (19D)	.79	.66	.60	.73	.55	.61	.74
12. Armor Crewman (19K)	.78	.56	.55	.71	.55	.42	.72
13. Military Police (31B)	.78	.64	.56	.68	.51	.63	.56
14. Human Resources Specialist (42A)	.83	.65	.64	.70	.41	.65	.71
15. Combat Medic Specialist (68W)	.84	.73	.63	.72	.62	.70	.51
16. Motor Transport Operator (88M)	.85	.73	.67	.76	.66	.66	.68
17. All MOS Combined	.81	.68	.62	.70	.59	.64	.60

Note. Army-wide PRS-S: $n = 13,001$ - $14,146$. MOS-specific PRS-S: 11B, $n = 2,225$ - $2,228$; 12B, $n = 399$ - 400 ; 13F, $n = 186$; 19D, $n = 275$ - 287 ; 19K, $n = 402$ - 403 ; 31B, $n = 1,666$ - $1,678$; 42A, $n = 385$; 68W, $n = 1,219$ - $1,457$; 88M, $n = 1,737$ - $1,806$. All MOS Combined, $n = 8,642$ - $8,952$. Ratings on PRS-S range from 1 and 5. PRS-S ratings from supervisors with a familiarity rating of 1 (“I have had little opportunity to observe this Soldier”) were excluded from analyses. Results based on fewer than 100 cases are not reported. All correlations are statistically significant ($p < .01$, one-tailed).

Table C.3. Correlations among the Supervisor Performance Rating Scales (PRS-S) in the In-Unit Validation Sample

Score/Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Overall Performance Composite														
2. Performing Core Warrior Tasks	.74													
3. Performing MOS-Specific Tasks	.75	.71												
4. Communicating with Others	.73	.59	.60											
5. Processing Information	.77	.63	.66	.68										
6. Solving Problems	.79	.66	.68	.64	.72									
7. Exhibiting Effort	.83	.63	.66	.61	.68	.71								
8. Exhibiting Personal Discipline	.80	.53	.53	.51	.58	.59	.65							
9. Contributing to the Team	.77	.50	.54	.52	.55	.57	.63	.67						
10. Physical Fitness & Bearing	.77	.48	.46	.43	.46	.49	.52	.54	.50					
11. Adjusting to the Army	.88	.60	.59	.55	.61	.63	.68	.67	.62	.65				
12. Following Safety Procedures	.71	.51	.54	.53	.56	.57	.58	.57	.57	.44	.57			
13. Developing Own Skills	.77	.55	.57	.52	.60	.62	.66	.61	.56	.52	.64	.57		
14. Managing Personal Matters	.71	.49	.49	.51	.54	.52	.53	.57	.54	.48	.58	.57	.54	
15. Leadership Potential	.77	.58	.55	.53	.59	.61	.64	.59	.55	.56	.75	.50	.62	.53

Note. Army-wide PRS-S, $n = 5,782-6,399$. Ratings on PRS range from 1 and 7. PRS ratings from supervisors 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.4. Correlations among the Peer Performance Rating Scales (PRS-P) in the IMT and In-Unit Validation Sample

Score/Scale	1	2	3	4	5	6	7	8	9	10	11
1. Overall Performance Composite		.86	.79	.72	.79	.84	.86	.31	.81	.82	.90
2. Effort and Discipline	.88		.70	.56	.63	.67	.70	.32	.67	.67	.76
3. Working Effectively with Other Soldiers	.85	.74		.46	.57	.64	.65	.31	.62	.58	.69
4. Physical Fitness	.74	.54	.50		.48	.50	.62	.20	.53	.53	.63
5. MOS Qualification Knowledge & Skill	.83	.69	.64	.56		.69	.63	.21	.53	.62	.71
6. Processing Information & Solving Problems							.71	.27	.63	.66	.72
7. Resilience and Adjustment	.88	.72	.70	.62	.70			.26	.64	.67	.76
8. Counterproductive Soldier Behaviors	.46	.47	.43	.25	.36		.38		.28	.19	.30
9. Going Above and Beyond	.86	.74	.73	.53	.65		.70	.39		.65	.69
10. Emergent Leadership											.68
11. Overall Performance Rating	.92	.79	.75	.67	.76		.79	.43	.75		

Note. Correlations below the diagonal reflect the IMT PRS-P, $n = 14,597$ - $14,784$. Correlations above the diagonal reflect the in-unit PRS-P, $n = 677$ - 705 . 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.5. Cross-instrument Correlations among Validation Analysis Criterion Scores in the IMT and In-Unit Validation Samples

Criterion	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Knowledge & Skill Composite		.94	.09					.01	.14		-.07		.00	.01	.02
2. Warrior Tasks and Battle Drills Job Knowledge Test	.86		.09					.01	.13		-.08		-.01	-.01	.01
3. Commitment & Fit Composite	.13	.14		.60	.44	-.25	.48	.08	.15	.11	-.15		.04	.02	.02
4. Retention Cognitions Composite	.00	.00	.52		.27	-.29	.30	.11	.10	.09	-.13		-.03	-.10	-.07
5. Peer Leadership Composite	.02	.00	.42	.23		-.08	.50	.21	.15	.17	-.12		-.01	-.03	.04
6. Counterproductive Soldier Behavior	-.18	-.19	-.33	-.17	-.03		-.28	-.10	-.06	-.11	.15		-.03	.03	-.02
7. Resilience	.07	.09	.54	.26	.48	-.29		.20	.18	.15	-.19		.04	-.03	-.02
8. Army Physical Fitness Test	.05	.06	.07	.03	.12	-.05	.13		.22	.21	-.09		.00	.02	.05
9. PRS-S Overall Performance Composite	.10	.05	.07	.03				.21		.34	-.23		.02	.04	.08
10. PRS-P Overall Performance Composite	.08	.07	.10	.05	.13	-.11	.16	.28			-.16		.06	.06	.11
11. Disciplinary Actions (Y/N)	-.03	-.02	-.07	-.04	-.06	.08	-.08	-.16	-.14	-.18			.00	.02	.04
12. Initial Military Training Restarts	-.01	-.01	-.02	.00	-.01	.01	-.01	-.04	-.01	-.04	.06				
13. 6-mo Attrition	-.01	.01	.02	.00	-.03	.02	-.01	-.01	-.04	-.04	.00	.02			
14. 12-mo Attrition	.01	.05	.07	.00	.01	-.01	.04	.01	-.04	.02	.00	.02			
15. 24-mo Attrition	.02	.05	.06	.00	.01	-.01	.01	-.01	-.04	.02	.00	.01			

Note. Correlations below the diagonal reflect the IMT criteria, $n = 5,769$ -618,852. Correlations above the diagonal reflect the in-unit criteria, $n = 480$ -617,418. Missing values reflect the scales that were not administered in either IMT or In-Unit. Correlations in bold are statistically significant, $p < .05$ (two-tailed).

APPENDIX D

ARMY LIFE QUESTIONNAIRE SCALES

Scale	Items	Version
Affective Commitment	I feel a strong sense of belonging to the Army.	IMT & IU
	I feel personally attached to the Army.	IMT & IU
	I doubt I could become as attached to another organization as I am to the Army.	IMT & IU
Army Fit	Life in the Army is worse than I expected before I joined the service. [REVERSE]	IMT & IU
	The Army is a good match for me.	IMT & IU
	I do not fit very well in the Army. [REVERSE]	IMT & IU
	The Army fulfills my needs. ^a	IMT & IU
Army Life Adjustment	During training, I have seriously questioned the wisdom of my decision to join the Army. [REVERSE]	IMT
	In general, I have struggled to meet the requirements of training. [REVERSE]	IMT
	Looking back, I was not prepared for the challenges of training in the Army. [REVERSE]	IMT
	At times I have felt ready to quit the Army. [REVERSE]	IMT
Career Intentions	How likely is it that you will make the Army a career?	IMT & IU
MOS Fit	Given my interests, I would be better off in another MOS. [REVERSE]	IMT & IU
	I like the work I do in my MOS.	IMT & IU
	I am the right person for the type of work that my MOS requires.	IMT & IU
	My MOS allows me to perform the kind of work that I want to do.	IMT & IU
	The tasks I am learning to perform in my MOS are a good fit for my work interests. ^b	IMT
	The tasks I perform in my MOS are a good fit for my work interests. ^b	IU
MOS Satisfaction	How satisfied are you with your opportunity to perform work you find interesting?	IU
	How satisfied are you with the amount of variety in your work?	IU
	How satisfied are you with your opportunity to use your aptitudes, experience, and training?	IU
	My MOS matches the expectations I had before joining. ^b	IU
	I would like to stay in my current MOS throughout my Army career. ^b	IU
Reenlistment Intentions	How likely is it that you will re-enlist in the Army? ^a	IMT & IU
Organizational Citizenship Behavior	Seek out a challenging assignment that is above and beyond your regular duties.	IMT & IU
	Go out of the way to encourage or praise another Soldier in training.	IMT & IU
	Take initiative to practice or develop skills outside of the classroom or programmed exercises.	IMT & IU
	Volunteer to help another Soldier practice or learn skills during, or outside of, programmed instruction.	IMT & IU
	Demonstrate concern about the image or reputation of your class.	IMT
	Demonstrate concern about the image or reputation of your unit.	IU

Scale	Items	Version
Leadership Behavior	Step in to deal with an interpersonal conflict before it escalates.	IMT & IU
	Keep your class focused on our goals and objectives.	IMT
	Monitor your environment for changes that may impact the mission during training exercises.	IMT
	Monitor your environment for changes that may impact the unit's mission.	IU
	Keep your unit focused on our goals and objectives.	IU
Motivation to Lead - Affective	I have a tendency to take charge in most groups or teams that I work in.	IMT & IU
	Most of the time, I prefer being a leader rather than a follower when working in a group.	IMT & IU
	I usually want to be the leader in the groups that I work in.	IMT & IU
Motivation to Lead - Noncalculative	I would want to know what's in it for me if I am going to agree to lead a group. [REVERSE]	IMT & IU
	I will never agree to lead if I cannot see any benefits from accepting that role. [REVERSE]	IMT & IU
	I would only agree to be a group leader if I know I can benefit from that role. [REVERSE]	IMT & IU
Motivation to Lead - Social-normative	I agree to lead whenever I am asked or nominated by the other members.	IMT & IU
	I was taught to believe in the value of leading others.	IMT & IU
	It is an honor and privilege to be asked to lead.	IMT & IU
Counterproductive Work Behavior	Try to keep your mistakes to yourself.	IMT & IU
	Leave a mess for someone else to clean up.	IMT & IU
	Speak poorly about the Army to others.	IMT & IU
	Take safety training or safety procedures less seriously.	IMT & IU
	Take an unauthorized break.	IMT & IU
	Openly question the decisions, intent, or wisdom of a supervisor.	IMT & IU
	Give another Soldier the "silent treatment."	IMT & IU
	Take supplies when not authorized to do so.	IMT & IU
	Waste time in training.	IMT
Resilience	Waste time on the job or in training.	IU
	I was confident in my ability to get through the stressful situation.	IMT & IU
	I liked to work out when I was stressed.	IMT & IU
	Knowing I had family or friends outside of the Army for support helped me deal with challenging situations.	IMT & IU
	Knowing I had support from my cadre and peers in the Army helped me deal with challenging situations.	IMT & IU
	When faced with a stressful situation in Army training, I am likely to come up with a strategy about what to do.	IMT
	When faced with a stressful situation in Army training, I am likely to remain calm.	IMT
	When faced with a stressful situation in Army training, I am likely to look for something good in what was happening.	IMT
	When faced with a stressful situation in my Army unit, I am likely to come up with a strategy about what to do.	IU
	When faced with a stressful situation in my Army unit, I am likely to remain calm.	IU
	When faced with a stressful situation in my Army unit, I am likely to look for something good in what was happening.	IU

Scale	Items	Version
Work-Life Balance	How satisfied are you with the support and concern the Army has for your family?	IMT & IU
	How satisfied are you with your amount of personal freedom?	IMT & IU
	How satisfied are you with the time available to pursue your personal life goals?	IMT & IU
	How satisfied are you with the amount of time you get to spend with family and friends?	IMT & IU
	How satisfied are you with your personal and family life?	IMT & IU
Deployment Satisfaction	Being deployed has had a positive effect on my Army career.	IU
	Since being deployed, I have second-guessed my decision to join the Army.	IU
	I have found my deployment to be an enriching experience.	IU
	Being deployed has increased the likelihood that I will re-enlist for another term.	IU

Note. IU = In-Unit.

^a Item was previously administered on the IMT version only but is now administered in both versions.

^b Item was added to the scale after the pilot study

APPENDIX E

SUMMARY OF BIVARIATE CORRELATIONS BETWEEN TAPAS SCALES AND SELECTED CRITERIA

Table E.1. Summary of the Bivariate Correlations between AFQT, TAPAS and Selected IMT Criteria

	Knowledge & Skill	WTBD JKT	Commitment & Fit	Retention Cognitions	Peer Leadership	CSB	Resilience
<i>n</i>	3,185 - 37,655	4,364 - 53,507	4,530 - 60,392	4,507 - 60,134	363 - 16,620	363 - 16,620	363 - 16,618
AFQT	.43	.41	-.04	-.11	-.05	-.02	-.02
<i>Individual TAPAS Scales</i>							
Achievement	.04	.03	.12	.07	.17	-.12	.15
Adjustment	.08	.07	.03	.00	-.05	-.04	-.02
Adventure Seeking ^a	.07	.09	.04	.00	—	—	—
Attention Seeking	-.01	.00	.04	.00	.12	.02	.07
Commitment to Serve ^a	-.04	-.04	.19	.21	.09	-.06	.10
Cooperation	-.03	-.02	.02	.00	-.02	-.08	.05
Courage ^a	.09	.09	.11	.06	.14	-.07	.13
Dominance	.04	.04	.10	.06	.27	-.05	.14
Even Tempered	.06	.05	.03	.02	-.01	-.09	.06
Intellectual Efficiency	.15	.14	.04	.03	.14	-.04	.08
Non-Delinquency	-.01	-.01	.05	.02	.00	-.13	.03
Optimism	.04	.04	.07	.01	.04	-.06	.10
Order	-.08	-.07	.01	.03	.07	-.07	.04
Physical Conditioning	.00	.00	.05	-.01	.11	-.01	.14
Responsibility ^a	.09	.08	.08	.02	.10	-.11	.11
Self-Control	.00	.00	.02	.03	.03	-.12	.02
Selflessness	-.02	-.01	.05	.05	.05	-.07	.08
Situational Awareness ^a	.03	.03	.06	.03	.09	-.07	.07
Sociability	-.10	-.08	.07	.05	.17	-.02	.10
Team Orientation ^a	-.04	-.04	.06	.03	.07	-.03	.08
Tolerance	.00	.00	.02	.05	.01	-.06	.06
<i>TAPAS Composites</i>							
Can-Do	.22	.19	.02	-.01	.04	-.02	.03
Will-Do	.03	.03	.12	.04	.22	-.08	.21
Adaptation	.09	.09	.02	-.03	-.02	.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 ($p < .01$, two-tailed).

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

Table E.1. (Continued)

		PRS-S: Overall Performance Composite	PRS-P: Overall Performance Composite	Disciplinary Incidents (Y/N)	Training Restarts
<i>n</i>	APFT 4,362 - 58,132	1,065 - 12,559	312 - 14,430	4,526 - 58,652	63,737 - 593,580
<i>AFQT</i>	.10	.06	.05	-.02	-.01
<i>Individual TAPAS Scales</i>					
Achievement	.10	.07	.08	-.07	-.01
Adjustment	.02	-.03	.02	-.02	.00
Adventure Seeking ^a	.09	-.01	-.04	-.03	-.01
Attention Seeking	.09	.02	.01	-.01	.00
Commitment to Serve ^a	-.02	-.04	-.02	-.03	.01
Cooperation	-.02	-.01	.03	-.02	.00
Courage ^a	.05	.00	.04	-.03	.00
Dominance	.12	.05	.06	-.05	-.01
Even Tempered	-.03	.00	.02	-.02	.01
Intellectual Efficiency	.04	.02	-.01	-.02	.01
Non-Delinquency	-.04	-.01	.02	-.03	.00
Optimism	.04	.03	.05	-.04	-.01
Order	.03	.01	.01	-.04	.00
Physical Conditioning	.26	.08	.13	-.09	-.03
Responsibility ^a	.03	.04	.08	-.04	-.01
Self-Control	-.01	.00	.11	-.02	.01
Selflessness	.01	.00	.01	.00	.00
Situational Awareness ^a	-.01	-.01	.03	-.04	.00
Sociability	.03	.00	-.02	.00	.01
Team Orientation ^a	.03	-.02	.05	-.03	.00
Tolerance	.00	-.01	-.02	.01	.02
<i>TAPAS Composites</i>					
Can-Do	.02	.02	.02	-.02	.00
Will-Do	.25	.10	.14	-.10	-.03
Adaptation	.17	.06	.12	-.06	-.02

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 ($p < .01$, two-tailed).

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

Table E.2. Summary of the Bivariate Correlations between AFQT, TAPAS and Selected In-Unit Criteria

	Knowledge & Skill	WTBD JKT	Commitment & Fit	Retention Cognitions	Peer Leadership	CSB	Resilience
<i>n</i>	1,031 - 6,684	1,031 - 6,677	1,099 - 7,734	131 - 877	132 - 878	132 - 878	132 - 878
<i>AFQT</i>	.44	.44	-.03	-.11	-.05	.10	-.02
<i>Individual TAPAS Scales</i>							
Achievement	.04	.04	.07	.01	.12	-.10	.07
Adjustment	.08	.09	.02	.06	-.01	-.01	.08
Adventure Seeking ^a	.10	.12	-.01	—	—	—	—
Attention Seeking	.00	.01	.00	.05	.07	.00	.01
Commitment to Serve ^a	-.04	-.04	.07	.15	.03	-.11	.10
Cooperation	-.04	-.03	.02	.10	-.04	-.03	-.04
Courage ^a	.07	.07	.04	-.01	-.01	.02	-.02
Dominance	.01	.01	.03	-.03	.18	-.03	.09
Even Tempered	.05	.05	.05	.08	-.03	-.12	.08
Intellectual Efficiency	.14	.14	.02	-.01	.07	.00	.05
Non-Delinquency	-.03	-.03	.06	.09	.00	-.14	.02
Optimism	.02	.01	.06	.02	.03	-.01	.13
Order	-.12	-.11	.00	-.01	.10	-.14	.04
Physical Conditioning	-.01	.00	.03	.03	.15	.04	.13
Responsibility ^a	.05	.06	.06	-.02	.05	-.08	.12
Self-Control	-.01	.00	.03	.00	-.02	-.06	-.02
Selflessness	-.04	-.04	.04	.03	.01	-.10	.06
Situational Awareness ^a	-.01	-.01	.02	.07	.16	-.04	.12
Sociability	-.11	-.12	.05	.00	.10	-.06	.07
Team Orientation ^a	-.08	-.06	.02	-.06	-.02	.05	-.02
Tolerance	-.03	-.04	.04	.00	-.05	-.05	-.02
<i>TAPAS Composites</i>							
Can-Do	.24	.24	.00	-.01	.01	.02	.02
Will-Do	.02	.02	.07	.02	.20	-.02	.16
Adaptation	.10	.10	.03	.06	.03	.06	.10

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 ($p < .01$, two-tailed).

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

Table E.2. (Continued)

		PRS-S: Overall Performance Composite	PRS-P: Overall Performance Composite	Disciplinary Incidents (Y/N)
<i>n</i>	APFT 1,079 - 7,516	884 - 6,016	131 - 692	1,114 - 7,857
<i>AFQT</i>	-.02	.08	.01	-.03
<i>Individual TAPAS Scales</i>				
Achievement	.07	.07	.07	-.07
Adjustment	.01	.02	-.12	-.01
Adventure Seeking ^a	.13	.05	—	.00
Attention Seeking	.08	.02	-.05	.02
Commitment to Serve ^a	.01	.02	-.08	-.01
Cooperation	-.04	.00	-.01	-.03
Courage ^a	.06	.02	.05	-.03
Dominance	.11	.04	.07	-.01
Even Tempered	-.03	.00	-.04	-.05
Intellectual Efficiency	.01	.02	-.02	-.01
Non-Delinquency	-.06	-.01	.03	-.06
Optimism	.02	.03	-.02	-.04
Order	.03	.00	.09	-.04
Physical Conditioning	.23	.07	.07	-.03
Responsibility ^a	.01	.06	.06	-.03
Self-Control	-.01	.02	—	-.06
Selflessness	-.02	.02	-.04	-.01
Situational Awareness ^a	.02	.03	.02	-.03
Sociability	.05	-.01	-.02	.01
Team Orientation ^a	.00	-.01	.02	.01
Tolerance	-.01	-.03	-.05	.00
<i>TAPAS Composites</i>				
Can-Do	-.02	.04	-.04	-.02
Will-Do	.22	.10	.09	-.06
Adaptation	.14	.06	.01	-.04

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 ($p < .01$, two-tailed).

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

Table E.3. Summary of the Bivariate Correlations between AFQT, TAPAS, and Selected Attrition Criteria (Regular Army Only)

<i>n</i>	Attrition		
	6-Mos	12-Mos	24-Mos
	33,412 - 320,593	33,141 - 285,695	32,428 - 229,487
<i>AFQT</i>	-.06	-.06	-.08
<i>Individual TAPAS Scales</i>			
Achievement	-.02	-.02	-.02
Adjustment	-.01	-.01	-.01
Adventure Seeking ^a	-.03	-.03	-.02
Attention Seeking	-.03	-.03	-.01
Commitment to Serve ^a	.01	.01	.02
Cooperation	.01	.01	.01
Courage ^a	.00	.00	.01
Dominance	-.03	-.03	-.02
Even Tempered	.00	.00	-.01
Intellectual Efficiency	-.01	.00	.00
Non-Delinquency	.02	.02	.01
Optimism	-.02	-.02	-.02
Order	.01	.01	.01
Physical Conditioning	-.07	-.08	-.07
Responsibility ^a	.00	.00	-.01
Self-Control	.00	.00	.00
Selflessness	.02	.02	.02
Situational Awareness ^a	.00	.00	.00
Sociability	.01	.01	.03
Team Orientation ^a	-.03	-.03	-.03
Tolerance	.01	.01	.01
<i>TAPAS Composites</i>			
Can-Do	-.01	-.01	-.02
Will-Do	-.07	-.07	-.06
Adaptation	-.07	-.07	-.07

Note. Correlations in bold are statistically significant ($p < .01$, two-tailed).

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