Technical Report 1267

Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a Non-Cognitive Screen for New Soldier Selection

Deirdre J. Knapp (Ed.)
Human Resources Research Organization

Tonia S. Heffner (Ed.)
U.S. Army Research Institute

July 2010

United States Army Research Institute
For the Behavioral and Social Sciences

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

A Directorate of the Department of the Army Deputy Chief of Staff, G1

Authorized and approved for distribution:

MICHELLE SAMS, Ph.D.
Director

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Brian Tate, U.S. Army Research Institute
Nehama Babin, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Technical Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-ZXM, 2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926.

FINAL DISPOSITION: This Technical Report may be destroyed when it is no longer needed. Please 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)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
July 2010	Final Report	July 2008 – March 2009

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a Non-Cognitive Screen for New Soldier Selection	DASW01-03-D-0015, DO #0049
	5b. PROGRAM ELEMENT NUMBER 622785
6. AUTHOR(S) Deirdre J. Knapp (Ed.) (Human Resources Research Organization), and Tonia S. Heffner (Ed.) (U.S. Army Research Institute)	5c. PROJECT NUMBER A790
	5d. TASK NUMBER 257
	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 2511 Jefferson Davis Highway Arlington, VA 22202-3926 Source Code: 597861	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Technical Report 1267

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Contracting Officer's Representative and Subject Matter POC: Tonia Heffner,

14. ABSTRACT *(Maximum 200 words)*:

The Army needs the best personnel available to meet the emerging demands of the 21st century. Accordingly, the Army is seeking recommendations on experimental non-cognitive predictor measures (e.g., interests, values, temperament) that could enhance entry-level Soldier selection and classification decisions. The U. S. Army Research Institute for the Behavioral and Social Sciences (ARI) is conducting a longitudinal criterion-related validation research effort to collect data to inform these recommendations. Experimental predictor measures of individual differences in temperament and job interests were administered at Army Reception Battalions to 8,103 new Soldiers. At the end of training, archival criterion data were collected for 7,599 Soldiers and supplemented with for-research-only criteria for 1,194 Soldiers. The results support the Tailored Adaptive Personality Assessment (TAPAS) and Work Preferences Assessment (WPA) as candidates for a new Soldier screen. Based on these results, the Army has implemented the TAPAS as an operational test for applicants and is pursing further research on the WPA. An operational test and evaluation (IOT&E) has been initiated to evaluate the new screen.

15. SUBJECT TERMS

Behavioral and social science, personnel, criterion-related validation, selection and classification, manpower

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	104	Ellen Kinzer Technical Publications Specialist (703) 602-8049

Standard Form 298

TECHNICAL REPORT 1267

Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a Non-Cognitive Screen for New Soldier Selection

Deirdre J. Knapp (Ed.)
Human Resources Research Organization

Tonia S. Heffner (Ed.)
U.S. Army Research Institute

Personnel Assessment Research Unit
Michael G. Rumsey, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

July 2010

Army Project Number
622785A790

Personnel, Performance
and Training Technology

Acknowledgements

There are individuals not listed as authors who have contributed significantly to the work described in this report. Dr. Richard Hoffman of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) provided additional oversight and support during the data collection efforts. Other ARI members who supported the data collections necessary for this effort include Drs. Nehama Babin, Elizabeth Brady, Kelly Ervin, Colanda Howard, Peter Schaefer, Teri Taylor, Michael Wesolak, Mark Young, and Mr. Doug Dressel.

We also acknowledge the support provided by LTG Michael Rochelle, Deputy Chief of Staff, Army G-1 and LTG Benjamin Freakley, Commanding General, U.S. Army Accessions Command. This work would not have been possible without the assistance from the thousands of Soldiers who participated in the research, the hundreds of non-commissioned officers and civilians who assisted with the data collection coordination, and the Army Test Program Advisory Team Members (ATPAT) who provided us with guidance throughout the course of this effort.

EXPANDED ENLISTMENT ELIGIBILITY METRICS (EEEM): RECOMMENDATIONS ON A NON-COGNITIVE SCREEN FOR NEW SOLDIER SELECTION

EXECUTIVE SUMMARY

Research Requirement:

In addition to educational, physical, and moral screens, the U.S. Army relies on a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), the Armed Forces Qualification Test (AFQT), to select new recruits into the Army. 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).

The *Future Force Performance Measures (Army Class)* project began in 2006 with contract support from the Human Resources Research Organization (HumRRO; Knapp & Heffner, 2009). This 6-year research effort includes both a concurrent and a longitudinal validation to investigate the selection and classification potential of a number of experimental non-cognitive predictors that might be used to supplement the ASVAB for pre-enlistment testing. After the Army Class research was underway, ARI initiated the *Expanded Enlistment Eligibility Metrics (EEEM)* project. EEEM goals are similar to Army Class, but the focus is specifically on Soldier selection (not classification) and the time horizon is much shorter. Specifically, EEEM requires selection of one or more promising new predictor tests and deriving a suitable screening algorithm based on the measure. The EEEM project capitalized on the existing Army Class data collection procedure and, thus, the EEEM sample was a subset of the Army Class data.

Procedure:

For the Army Class project, experimental predictors were administered to roughly 11,000 new Soldiers representing all Components (Regular Army, Reserve, National Guard) in 2007 and early 2008. The experimental predictors were administered to new Soldiers as they entered the Army through one of four reception battalions. The predictor measures included (a) three temperament measures (Assessment of Individual Motivation [AIM], Tailored Adaptive Personality Assessment System [TAPAS], and Rational Biodata Inventory [RBI]), (b) a predictor situational judgment test (PSJT), and (c) two person-environment (P-E) fit measures (Work Preferences Assessment [WPA] and Army Knowledge Assessment [AKA]). In addition, we obtained ASVAB scores, including the Assembling Objects (AO) test, a spatial ability measure. AO is currently administered with the ASVAB but has limited use for Army personnel decisions.

The criterion measures include archival data from Army records and for-research-only measures administered to Soldiers in six job-specific samples at the end of training. For all Soldiers, we obtained the available data on attrition (through the first 6 months of service) and performance during training from administrative records. The for-research-only criterion measures

administered were (a) a military occupational specialty (MOS)-specific job knowledge test (JKT), (b) MOS-specific and Army-wide performance ratings collected from training instructors and peers, and (c) a questionnaire measuring Soldiers' experiences and attitudes towards the Army through training (the Army Life Questionnaire [ALQ]).

Four factors were considered in evaluating which experimental predictor measures represented the "best bets" for enhancing new Soldier selection: (a) incremental validity (over the AFQT), (b) subgroup differences, (c) susceptibility to faking and coaching effects, and (d) administration time.

Findings:

Based on the results of this research, the TAPAS appears to be the "best bet" predictor measure for enhancing new Soldier selection in an operational setting. It exhibited high incremental validity, few subgroup differences, has the potential to be less susceptible to faking or coaching effects than other candidate measures (e.g., AIM, RBI), is based on a scientifically advanced item response theory psychometric model, and can be administered in a reasonable timeframe. Additional analyses were conducted to develop a specific recommendation for how the TAPAS should be used to screen Army applicants.

Utilization and Dissemination of Findings:

These finding have been disseminated to Army Senior Leaders. Based on their recommendations, this research provided a foundation for administration of selected experimental predictor measures to new Army applicants in an operational setting, as part of an initial operational test and evaluation (IOT&E) starting in fall 2009.

EXPANDED ENLISTMENT ELIGIBILITY METRICS (EEEM): RECOMMENDATIONS ON A NON-COGNITIVE SCREEN FOR NEW SOLDIER SELECTION

CONTENTS

Page

APPENDICES

Page

LIST OF TABLES

Page

Page

Page

EXPANDED ENLISTMENT ELIGIBILITY METRICS (EEEM): RECOMMENDATIONS ON A NON-COGNITIVE PREDICTOR SCREEN FOR NEW SOLDIER SELECTION

CHAPTER 1: INTRODUCTION

Deirdre J. Knapp, Michael J. Ingerick (HumRRO),
and Tonia S. Heffner (ARI)

Background

The Personnel Assessment Research Unit (PARU) of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is responsible for conducting manpower and personnel research for the Army. The focus of PARU's research is maximizing the potential of the individual Soldier through maximally effective selection, classification, and retention strategies.

In addition to educational, physical, and moral screens, the U.S. Army has relied on a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), the Armed Forces Qualification Test (AFQT), to select new recruits into the Army. 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).

The *Validating Future Force Performance Measures*, also known as the Army Class project, began in 2006 with contract support from the Human Resources Research Organization (HumRRO; Knapp & Heffner, 2009). This 6-year research effort includes both a concurrent and a longitudinal validation to investigate the selection and classification potential of a number of experimental non-cognitive individual difference measures that might be used to supplement the ASVAB for pre-enlistment testing. Experimental predictors were administered to new Soldiers in 2007 and early 2008. Since then, Army Class researchers have obtained attrition data from Army records and collected training criterion data on a subset of the Soldier sample. Job performance criterion data were collected from Soldiers in the Army Class longitudinal validation sample in 2009 and a second round of job performance data will be collected in 2010.

After the Army Class research was underway, ARI initiated the *Expanded Enlistment Eligibility Metrics (EEEM)* project. EEEM goals were similar to Army Class, but the focus was specifically on Soldier selection (not classification) and the time horizon is much shorter. Specifically, EEEM required selection of one or more promising new predictor measures and deriving a suitable screening algorithm based on the measure or measures. The EEEM project capitalized on the existing Army Class data collection procedure and, thus, the EEEM sample was a subset of the Army Class data.

Although training criterion data for the Army Class project were collected through the course of 2008, the EEEM data analyses needed to be completed earlier so that they could include data from selected military occupational specialties (MOS) which would be collected through May 2008, before the final data collection of Army Class. An Army Class technical

report details the predictor and training criterion data collections and reports results of analyses based on the full sample (Knapp & Heffner, 2009). The present report summarizes the research design of the Army Class/EEEM longitudinal validation, with particular focus on two predictor measures that were added to support the requirements of the EEEM initiative. We also present analyses that factored into recommendations for an initial operational test and evaluation (IOT&E) of a promising new predictor screen based on data collected through the middle of 2008. Supporting analyses of the Army Class dataset are provided in Knapp and Heffner (2009).

Current Research Questions

The EEEM research summarized in this report was conducted to answer two questions:

- Which experimental non-cognitive predictor measures represent the "best bets" for enhancing the recruitment and screening of new applicants into the Army over the existing AFQT, focusing specifically on high school diploma graduates with lower AFQT scores?

- How can the Army best use selected experimental predictor measures to screen applicants?

To answer these questions, the primary goal is to identify predictor measure(s) and a performance screen that demonstrates the greatest potential to maximize outcomes valued by the Army. The performance screen is intended to distinguish between applicants who will and will not perform well in the Army based on their responses to a set of assessments. Performance is a broadly defined construct that includes "can do" outcomes such as job knowledge tests scores and course grades, and "will do," or motivational, outcomes such as physical fitness, training course completion, attrition, and attitudes. At the time this research was initiated, the outcomes were specifically the training performance and retention of new Soldiers who are high school diploma graduates (Educational Tier 1). The Army is currently using the Tier Two Attrition Screen (TTAS) to screen applicants who are not high school diploma graduates but yet have a lower risk of attriting. TTAS includes a non-cognitive measure (the Assessment of Individual Motivation [AIM]) as a primary component of the screen. However, no such measures are presently being used to screen applicants who are high school diploma graduates. Accordingly, the EEEM analyses focused on Educational Tier 1 Soldiers.

Overview of Report

Chapter 2 describes the Army Class/EEEM longitudinal validation research design, including the data collection samples and measures. A more comprehensive description is provided in Knapp and Heffner (2009). Chapter 3 provides detailed information about the Tailored Adaptive Personality Assessment System (TAPAS) measure that was added to the research plan to support the EEEM initiative. Chapter 4 briefly summarizes information about the other predictor measures, including their psychometric properties in the EEEM sample. Chapter 5 presents the findings of analyses directed at answering the key EEEM research questions. The report concludes with Chapter 6, which discusses the way forward given the results of this work and subsequent evaluation and consideration by Army policy-makers.

CHAPTER 2: LONGITUDINAL RESEARCH DESIGN

Deirdre J. Knapp (HumRRO), [1]
Tonia S. Heffner, and Kimberly S. Owens (ARI)

This chapter describes the research design for the Army Class/EEEM longitudinal validation, beginning with the sample selection strategy and plan for collecting data from participating Soldiers at up to four points in time. Selection, development, and descriptions of the training criterion measures and then the experimental pre-enlistment measures that were selected to predict the criteria are presented.

Data Collection Points and Sample

In 2007 through early 2008, predictor data were collected from new Soldiers as they entered the Army through one of four Army reception battalions. Soldiers in the longitudinal predictor data collection were drawn from two types of samples: (a) MOS-specific samples targeting six entry-level jobs and (b) an Army-wide sample with no MOS-specific membership requirements. The six targeted MOS were:

- 11B (Infantryman)
- 19K (Armor Crewman)
- 31B (Military Police)
- 63B (Light Wheel Vehicle Mechanic)
- 68W (Health Care Specialist)
- 88M (Motor Transport Operator)

Criterion Measures

Training performance criterion data were subsequently obtained on participating Soldiers at the completion of their Initial Entry Training (IET)—either Advanced Individual Training (AIT) or One-Station Unit Training (OSUT), as applicable to the MOS. This criterion data collection included only Soldiers who were in one of the six MOS-specific samples described above. Additional training performance and attrition criterion data were drawn from archival records including records from Soldiers who were in the Army-wide sample but were not assigned to one of the six target MOS.

Table 2.1 lists all of the training criterion measures administered in the Army Class/EEEM longitudinal validation. Full descriptions of these measures are provided in Knapp and Heffner (2009) and development of the job knowledge tests, rating scales, and Army Life Questionnaire (ALQ) is detailed in Moriarty, Campbell, Heffner, and Knapp (2009). Moriarty et al. (2009) also provides copies of the rating scales and ALQ.

[1] Most of this chapter is drawn from the companion Army Class project report (Knapp & Heffner, 2009).

Table 2.1. Summary of Longitudinal Validation Training Criterion Measures

Criterion Measure	Description
Computer-Administered (MOS-specific sample)	
MOS-Specific Job Knowledge Test (JKT)	Measures Soldiers' knowledge of the basic facts, principles, and procedures required in MOS training (e.g., the major steps in loading a tank main gun, the main components of an engine). Each JKT consists of about 70 items representing a mix of item formats (e.g., multiple-choice, multiple-response, rank order, and drag and drop).
MOS-Specific and Army-Wide (AW) Performance Rating Scales (PRS)	Measures Soldiers' performance during AIT/OSUT on two categories of dimensions: (a) MOS-specific (e.g., performs preventive maintenance checks and services, troubleshoots vehicle and equipment problems) and (b) Army-wide (e.g., exhibits effort, supports peers, demonstrates physical fitness). The PRS were designed to be completed by the supervisors and peers of the Soldier being rated.
Army Life Questionnaire (ALQ)	Measures Soldiers' self-reported attitudes and experiences through the end of AIT/OSUT. The training ALQ consists of 13 scales. The content of the 13 scales covers two general categories: (a) commitment and other retention-related attitudes towards the Army and MOS at the end of AIT/OSUT (e.g., perceived fit with Army; perceived fit with MOS) and (b) performance and adjustment during IET (e.g., adjustment to Army life, number of disciplinary incidents during IET).
Archival (Army-wide)	
Attrition	Attrition data were obtained on participating Regular Army Soldiers through their first 6 months of service in the Army. These data were extracted from the Tier Two Attrition Screen (TTAS) database maintained by the U.S. Army Accessions Command.
Initial Entry Training (IET) Performance and Completion	Operational IET performance and completion data were obtained from two Army administrative personnel databases: (a) Army Training Requirements and Resources System (ATRRS) and (b) Resident Individual Training Management System (RITMS). Soldier data on three IET-related criteria were extracted from these databases: (a) graduation from AIT/OSUT; (b) number of times recycled through AIT/OSUT; and (c) average AIT/OSUT exam grade.

MOS-Specific Job Knowledge Tests

Most of the training job knowledge test (JKT) items are in a multiple-choice format with two to four response options. However, other formats, such as multiple response (i.e., check all that apply), rank ordering, and matching are also used. The number of items on the six training JKTs range from 60 to 82. The items make liberal use of visual images to make them more realistic and to reduce reading requirements for the test. The JKTs yield a single overall score.

Performance Rating Scales

The training-oriented Army-wide rating scales measure aspects of Soldier performance critical to all Soldiers, such as the amount of effort they exhibit, commitment to the Army, and personal discipline. We used a relatively atypical bipolar format for these scales (see example in Figure 2.1). Seven of the eight dimensions had multiple rating scales and there was a single rating of "MOS Qualification and Skill") for a total of 21 individual ratings. Each response scale

has a behavioral statement on the low end (rating of 1) and on the high end (rating of 5). Five of the Army-wide dimensions were used in the EEEM analyses[2]. They are defined in Table 2.2.

C. Personal Discipline						
Behaves consistently with Army Core Values; demonstrates respect in word and actions towards superiors, instructors, and others; adheres to training behavior limitations (for example, use of cell phones and tobacco).						
Complains about requirements and directions; may delay or resist following directions.	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	Follows requirements and directions willingly.

Figure 2.1. Example Army-wide training rating scale.

Table 2.2. Army-Wide Performance Rating Scales (PRS) Used in EEEM Analyses

Dimension	Description
Effort	Three-scale measure assessing Soldiers' persistence and initiative demonstrated when completing study, practice, preparation, and participation activities during AIT/OSUT (e.g., persisting with tasks, even when problems arose; paying attention in class and studying hard).
Physical Fitness and Bearing	Three-scale measure assessing Soldier's physical fitness and effort exhibited to maintain self and appearance to standards (e.g., meeting or exceeding basic standards for physical fitness, dressing and carrying self according to standard).
Personal Discipline	Five-scale measure assessing Soldier's willingness to follow directions and regulations and to behave in a manner consistent with the Army's Core Values (e.g., showing up on time for formations, classes, and assignments; showing proper respect for superiors).
Support for Peers	Three-scale measure assessing Soldier's support for and willingness to help their peers (e.g., offering assistance to peers that are ill, distressed, or failing behind; treating peers with respect, regardless of cultural, racial, or other differences).
Peer Leadership	Three-scale measure assessing Soldier's proficiency in leading their peers when assigned to an AIT/OSUT leadership position, (e.g., gaining the cooperation of peers; taking on leader roles as assigned; giving clear directions to peers).

Note. Three scales were omitted from the EEEM analyses: Commitment and Adjustment to the Army, Common Warrior Tasks Knowledge and Skill, MOS Qualification Knowledge and Skill.

We used a more traditional format for the MOS-specific rating scales. As shown in the example in Figure 2.2, each rating scale measures a single aspect of MOS-specific performance and is rated on a 7-point response scale. Multiple bulleted summary statements anchor the low, middle, and high ends of the scale. The number of dimensions varies depending on the MOS, but ranges from five to eight. The EEEM analyses used a single composite score from these scales.

[2] A subset of the criterion scores were used in the validation analyses. An explanation for how they were selected is provided in Chapter 5.

A. Learns to Use Aiming Devices and Night Vision Devices How well has the Soldier learned to engage targets with aiming devices, to zero sights, and to operate and maintain night vision devices?		
1 2	3 4 5	6 7
– Is unable to engage targets with bore light and other aiming devices. – Cannot zero sights accurately, in daylight or at night; does not understand field zero.	– Is able to engage targets with bore light and other aiming devices with practice and coaching. – Zeroes sights accurately, but not quickly, both in daylight and at night; can apply field zero.	– Is extremely proficient in engaging targets with all types of aiming devices. – Zeroes sights quickly and accurately without assistance both in daylight and at night; applies field and expedient zero methods.

Figure 2.2. Example MOS-specific training criterion rating scale.

Army Life Questionnaire (ALQ)

The ALQ was designed to measure Soldiers' self-reported attitudes and experiences through the end of IET. The training ALQ consists of 13 scales, the content of which falls into two general categories: (a) commitment and other retention-related attitudes towards the Army and MOS (e.g., perceived fit with Army; perceived fit with MOS) and (b) performance and adjustment (e.g., adjustment to Army life, number of disciplinary incidents during IET). Six of the 13 ALQ scales were selected for use in the EEEM analyses. They are defined in Table 2.3.

Table 2.3. Description of the Training Army Life Questionnaire Scales Used in EEEM Analyses

Scale	Description
Commitment and Retention-Related Attitudes	
Attrition Cognitions	Four-item scale measuring the degree to which Soldiers think about attriting before the end of their first-term (e.g., "How likely is it that you will complete your current term of service?").
Career Intentions	Five-item scale measuring Soldiers' intentions to re-enlist and to make the Army a career (e.g., "How likely is it that you will re-enlist in the Army?").
Army Fit	Six-item scale measuring Soldiers' perceived fit with the Army in general (e.g., "The Army is a good match for me.").
Affective Commitment	Seven-item scale measuring Soldiers' emotional attachment to the Army (e.g., "I feel like I am part of the Army 'family.' ").
Initial Entry Training (IET) Performance and Adjustment	
Disciplinary Incidents	Two-item measure (each item is segmented into multiple sub-questions) that asks Soldiers to self-report whether they had been involved in a series of disciplinary incidents (e.g., "While in the Army, have you ever been formally counseled for lack of effort?").
Army Physical Fitness Test (APFT) Score	Single-item asking Soldiers to self-report their most recent APFT score.

Note. The scales not included in the EEEM analyses included MOS Fit, Normative Commitment, Adjustment to Army Life, Number of IET Achievements, Number of IET Failures, Self-Rated AIT/OSUT Performance, and Self-Ranked AIT/OSUT Performance.

Archival Criterion Data

Attrition

Attrition data were obtained on participating Soldiers through their first 6 months of service in the Army. The 6-month timeframe was selected because (a) it roughly corresponds to the completion of IET for most Soldiers in most MOS and (b) it balanced the maturity of the attrition criterion (i.e., longer timeframes lead to more stable estimates) with the number of Soldiers on whom attrition data were available at the time analyses were conducted. Attrition information was extracted for participating Soldiers from the TTAS database maintained by the U.S. Army Accessions Command.

IET Performance and Completion

IET performance and completion data were obtained from two administrative personnel databases: (a) Army Training Requirements and Resources Systems (ATRRS) and (b) Resident Individual Training Management System (RITMS). Soldier data on three IET-related criteria were constructed from data extracted from these databases: (a) graduation from AIT/OSUT, (b) number of times recycled through AIT/OSUT, and (c) average AIT/OSUT exam grade.

Predictor Measures

Table 2.4 summarizes the predictor measures selected for inclusion in the joint Army Class/EEEM research. Brief descriptions of the measures are provided in this section. Because it is a particular focus in the EEEM research, more detailed information about the TAPAS is provided in Chapter 3.

Table 2.4. Summary of Longitudinal Validation Predictor Measures

Predictor Measure	Description
Baseline Predictor	
Armed Forces Qualification Test (AFQT)	Measures general cognitive ability. The AFQT is a unit-weighted composite based on four Armed Services Vocational Aptitude Battery (ASVAB) subtests (Arithmetic Reasoning, Mathematics Knowledge, Word Knowledge, and Paragraph Comprehension). Applicants must meet a minimum score on the AFQT to enter the Army.
Cognitive Predictor	
Assembling Objects (AO)	Measures spatial ability. AO is currently administered as part of the ASVAB, but until recently had not been used by the Army to screen or select applicants. AO is now included in the Two Tier Attrition Screen (TTAS) to screen applicants who have not earned a high school degree.
Temperament Predictors	
Assessment of Individual Motivation (AIM)	Measures six temperament characteristics predictive of first-term Soldier attrition and performance (e.g., work orientation, dependability, adjustment). Each item consists of four behavioral statements. Respondents are asked to indicate which statement is most descriptive of themselves and which statement is least descriptive of themselves

Table 2.4. Summary of Longitudinal Validation Predictor Measures (cont'd)

Predictor Measure	Description
Temperament Predictors (Continued)	
Tailored Adaptive Personality Assessment System (TAPAS-95s)	Measures 12 dimensions or temperament characteristics predictive of first-term attrition and performance (e.g., dominance, attention-seeking, intellectual efficiency, physical conditioning). Uses a multidimensional pairwise preference (MDPP) format in which respondents indicate which of two statements is most like them.
Rational Biodata Inventory (RBI)	Measures temperament and motivational characteristics important to entry-level Soldier performance and retention. Items ask respondents about their past behavior, experiences, and reactions to previous life events (e.g., the extent to which they enjoyed thinking about the "plusses and minuses" of alternative approaches to solving a problem).
Predictor Situational Judgment Test (PSJT)	Measures respondents' judgment and decision-making proficiency across situations commonly encountered prior to or during the first enlistment term (e.g., dealing with a difficult co-worker). Each item consists of a description of a problem situation and a list of four alternative actions that the respondent might take in that situation. Respondents rate the effectiveness of each action.
Person-Environment (P-E) Fit Predictors	
Work Preferences Assessment (WPA)	Measures respondents' preferences for different kinds of work activities and settings offered by different jobs (e.g., working with others, repairing machines or equipment). Items ask respondents to rate how important a series of characteristics are to their ideal job. Content is based on Holland's (1997) theory of vocational personality and work environment.
Army Knowledge Assessment (AKA)	Measures respondents' understanding or expectations about the kinds of work activities and settings typically offered by the Army. Respondents are asked to read a brief description of six work settings and then rate the extent to which they think each setting describes the Army. Like the WPA, content is based on Holland's (1997) theory of vocational personality and work environment.

Armed Forces Qualification Test (AFQT)

The AFQT is a unit-weighted composite of four ASVAB tests (Arithmetic Reasoning, Math Knowledge, Word Knowledge, and Paragraph Comprehension). Scores on the AFQT reflect an applicant's general cognitive aptitude and are one of the metrics, in addition to applicant's high school degree status, used to judge recruit potential. Examinees are classified into categories based on their AFQT percentile scores (Category I = 93-99, Category II = 65-92, Category IIIA = 50-54, Category IIIB = 31-49, Category IV = 10-30, Category V = 1-9). The AFQT served as the baseline against which the experimental predictors were evaluated.

Assembling Objects (AO)

AO is an ASVAB subtest that measures spatial ability and was first developed in Project A (Russell et al., 2001). The items are graphical in nature, requiring respondents to visualize how an object will look when its parts are put together correctly. AO is included in the TTAS enlistment screen for applicants who do not have a high school diploma.

Assessment of Individual Motivation (AIM)

AIM was added to the Army Class longitudinal validation as part of the EEEM initiative. The AIM was developed to improve and advance the promising Assessment of Background and Life Experiences (ABLE) developed in Project A (White & Young, 1998; White, Young, & Rumsey, 2001). The AIM, which measures six temperament characteristics predictive of first-term Soldier attrition and performance, uses a forced-choice format to reduce fakability and improve the accuracy of the self-report information (see Table 2.5). Each item consists of four behavioral statements (i.e., tetrads). Respondents are asked to indicate which statements are most and least descriptive of themselves. The version of AIM administered in this research has 30 items. An alternate form of the AIM is used operationally by the Army to screen applicants who are not high school diploma graduates into the Army's TTAS program.

Table 2.5. Description of AIM Dimensions

Scale	Description
Work Orientation	The tendency to strive for excellence in the completion of work-related tasks. Persons high on this construct seek challenging work activities and set high standards for themselves. They consistently work hard to meet these high standards.
Adjustment	The tendency to have a uniformly positive affect. Persons high on this construct maintain a positive outlook on life, are free of excessive fears and worries, and have a feeling of self-control. They maintain their positive affect and self-control even faced with stressful situations.
Agreeableness	The tendency to interact with others in a pleasant manner. Persons high on this construct get along and work well with others. They show kindness, while avoiding arguments and negative emotional outbursts directed at others.
Dependability	The tendency to respect and obey rules, regulations, and authority figures. Persons high on this construct are more likely to stay out of trouble in the workplace and avoid getting into difficulties with law enforcement officials.
Leadership	The tendency to seek out and enjoy being in leadership positions. Persons high on this scale are confident of their abilities and gravitate towards leadership roles in groups. They feel comfortable directing the activities of other people and are looked to for direction when group decisions have to be made.
Physical Conditioning	The tendency to seek out and participate in physically demanding activities. Persons high on this construct routinely participate in vigorous sports of exercise, and enjoy hard physical work.

Rational Biodata Inventory (RBI)

The RBI measures multiple temperament or motivational characteristics important to entry-level Soldier performance and retention (Kilcullen, Putka, McCloy, & Van Iddekinge, 2005). Items on the RBI ask respondents about their past behavior, experiences, and reactions to previous life events (e.g., the extent to which they enjoyed thinking about the plusses and minuses of alternative approaches to solving a problem, how frequently did they engage in physical activities) using multiple Likert-style response scales. The RBI yields scores on a range of attributes (e.g., Achievement Motivation, Cognitive Flexibility, Fitness Motivation, Hostility to Authority, Peer Leadership, Self-Efficacy, and Stress Tolerance). The RBI used in the Army

Class longitudinal validation has 101 items covering 14 attributes and is the same version used in the Select21 research which also examined new Soldier selection (Kilcullen et al., 2005).

Predictor Situational Judgment Test (PSJT)

The PSJT is a 20-item paper-and-pencil measure designed to assess an individual's judgment and decision-making proficiency in challenging situations (e.g., working with uncooperative peers to accomplish a task; determining when to handle a problem alone versus consulting a supervisor; Waugh & Russell, 2005). The PSJT targets five kinds of situations or dimensions important to first-term Soldier performance: (a) adaptability to changing conditions, (b) relating to and supporting peers, (c) teamwork, (d) self-management, and (e) self-directed learning. Each item consists of a description of a situation followed by four actions that might be taken in that situation. Respondents rate the effectiveness of each action on a 1 to 7 scale (from *Ineffective* to *Very Effective*). Although the PSJT items were written to reflect these dimensions, it is designed to yield a single total score.

Work Preferences Assessment (WPA)

The WPA is designed to assess an individual's preferences (or fit) for different kinds of work activities and environments (Van Iddekinge, Putka, & Sager, 2005). The 72 items comprising the WPA were written to measure each of the six dimensions and their subfacets underlying Holland's (1997) theory of vocational personality and work environment. According to Holland's theory, work interests are expressions of personality that can be used to categorize individuals and work environments into six types (or dimensions): Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C). For each dimension or facet, the WPA contains three types of items: (a) interests in work activities (e.g., "A job that requires me to teach others"), (b) interests in work environments or settings (e.g., "A job that requires me to work outdoors"), and (c) interests in learning opportunities (e.g., "A job in which I can learn how to lead others"). Respondents are asked to rate each item in terms of its importance to their ideal job using a 5-point Likert-type scale (1 = *Extremely unimportant to have in my ideal job* to 5 = *Extremely important to have in my ideal job*; Putka & Van Iddekinge, 2007).

The WPA yields six dimension scores (corresponding to each of the six RIASEC dimensions) and 14 facet scores (corresponding to facets underlying the six RIASEC dimensions). These raw scores can then be combined or modified based on additional data to obtain multiple, alternative sets of scores for use in one or more of the Army's personnel management objectives.

Army Knowledge Assessment (AKA)

The AKA is a 30-item instrument that assesses Soldiers' knowledge about the extent to which the current Army (in general) supports each RIASEC dimension (Van Iddekinge et al., 2005). Respondents read a brief description of six work settings and then rate the extent to which they think each setting describes the Army (1 = *Strongly Disagree* to 5 = *Strongly Agree*) . The AKA yields six dimension scores that correspond to the six RIASEC dimensions defined by Holland (1997). Conceptually, the AKA is distinguished from the WPA in that it asks whether respondents have realistic expectations about the interests that would be satisfied with Army life whereas the WPA asks whether respondents are interested in what Army life offers. Both are strategies for predicting person-environment fit.

Predictor and Criterion Data Collections

Predictor data were collected from new Soldiers entering four reception battalions during the period of May 2007 through February 2008, ensuring that the resulting sample would reflect the recruit variations anticipated over the course of a year. Data collection visits were scheduled with each reception battalion to optimize our ability to gather data on Soldiers in the six target MOS as well as to maximize the total number of Soldiers tested. Data were collected over the course of 31 data collection site visits. The test sessions took 2 to 2.5 hours. Participating Soldiers represented all components: Regular Army (RA), U.S. Army Reserve (USAR), and the Army National Guard (ARNG).

We collected criterion data as Soldiers in the Army Class/EEEM longitudinal validation target MOS completed AIT or OSUT. The training data collection schedule was driven by the flow of Soldiers in the predictor data collections and the length of training for each MOS. Most test sessions ran about 2 hours. The EEEM data collections were conducted from mid-September 2007 through early May 2008, although the Army Class criterion data collection continued through August 2008. Therefore, the EEEM sample is a subset of the Army Class sample.

Sample Descriptions

Predictor Sample

Predictor data were collected on over 11,000 new Soldiers. The full Army Class predictor sample included 10,814 Soldiers after the data was cleaned (e.g., excluding Soldiers missing more than 10% of the responses on a measure). Given the goals of the EEEM analyses as described in Chapter 1, the EEEM analysis sample was restricted to non-prior service Educational Tier 1 Soldiers. Table 2.6 summarizes the demographic characteristics and entry qualifications of the EEEM analysis sample.

Table 2.6. Descriptive Statistics for EEEM Predictor Analysis Sample

	MOS						Totals		
Subgroup	11B/C/X	19K	31B	63B	68W	88M	AW	*n*	%
Gender									
Male	1,182	375	960	274	152	270	3,083	6,296	77.7
Female	0	0	268	36	104	118	1,256	1,782	22.0
Race									
White	1,021	317	1,057	250	207	273	3,057	6,182	76.3
Black	67	27	86	41	29	88	921	1,259	15.5
Other	92	30	83	17	20	25	369	636	7.8
Ethnicity									
White Non-Hispanic	906	288	964	228	193	256	2,632	5,467	67.5
Hispanic	180	44	173	38	32	38	694	1,199	14.8
AFQT Category									
I-II	405	109	404	82	196	103	1,546	2,845	35.1
IIIA	286	92	361	70	58	75	981	1,923	23.7
IIIB	438	151	449	127	2	161	1,573	2,901	35.8
IV	50	23	11	30	0	46	227	387	4.8[a]
Totals	1,183	376	1,230	310	256	388	4,360	8,103	

Note. The figures reported do not add up to the totals due to missing data. Soldiers in this sample are non-prior military service in Educational Tier 1.

[a]This number exceeds the yearly percentage accessed into the Army because we intentionally oversampled the Category IV Soldiers.

The EEEM portion of the effort was initiated after the Army Class predictor data collection was underway and resulted in the addition of the AIM and TAPAS to the test administration plan mid-stream. Because of limited administration time, changes were made to the administration plan to ensure that data on the AIM and TAPAS were collected from a sufficient number of Soldiers. These changes involved temporarily suspending administration of the PSJT and, at one data collection site, rotating the measures in the instrument set so that each Soldier did not take one of predictor measures. As a result, sample sizes vary across predictor measures considerably, as shown in Table 2.7.

Table 2.7. Analysis Sample Sizes by Predictor Measure

Predictor Measure	*n*
Armed Forces Qualification Test (AFQT)	8,056
Assembling Objects (AO)	7,300
Tailored Adaptive Personality Assessment System (TAPAS-95s)	3,381
Assessment of Individual Motivation (AIM)	3,286 – 3,376
Rational Biodata Inventory (RBI)	6,517 – 6,518
Army Knowledge Assessment (AKA)	7,610 – 7,613
Work Preferences Assessment (WPA)	7,511 – 7,512
Predictor Situational Judgment Test (PSJT)	3,996

Note. Ranges reflect the fact that not all Soldiers had non-missing or valid scale scores for an entire measure.

Training Criterion Samples

There are two training criterion samples. The first comprises Soldiers from the target MOS who were administered the for-research-only criterion measures. Although there were six such MOS, only four – 11B, 19K, 31B, and 63B – had sample sizes sufficient for the EEEM analyses. The second comprises Soldiers from the entire non-prior service Educational Tier 1 predictor sample for which we were able to retrieve criterion data from archival records. Tables 2.8 and 2.9 describe the criterion sample completing the for-research-only training criterion measures. Specifically, Table 2.8 describes the sample by MOS and component; Table 2.9 describes the demographics of the sample by MOS. Comparable information is provided for the archival criterion sample in Tables 2.10 and 2.11.

Table 2.8. EEEM For-Research-Only Training Criterion Sample by MOS and Component

	Component			
MOS	RA	ARNG	USAR	Totals
11B	261	46	0	308
19K	188	64	0	254
31B	227	203	103	533
63B	41	43	15	99
Totals	718	356	119	1,194

Note. Three Soldiers were missing component information. The figures reported do not add up to the totals due to missing data. Soldiers in this sample were non-prior military service in Educational Tier 1.

Table 2.9. EEEM For-Research-Only Training Criterion Sample by MOS and Demographic Subgroup

	MOS				Subgroup Totals	
Subgroup	11B	19K	31B	63B	*n*	*%*
Gender						
Male	308	254	434	91	1,087	91.0
Female	0	0	98	8	106	8.9
Race						
White	270	221	466	83	1,040	87.1
Black	17	13	35	11	76	6.4
Other	19	18	31	5	73	6.1
Ethnicity						
White Non-Hispanic	233	203	414	72	922	77.2
Hispanic	50	26	84	14	174	14.6
AFQT Category						
I-II	104	79	166	27	376	31.5
IIIA	80	66	163	18	327	27.4
IIIB	111	104	200	45	460	38.5
IV	13	5	3	9	30	2.5
Totals	308	254	533	99	1,194	

Note. The figures reported by subgroup and MOS do not add up to the totals due to missing data. Soldiers indicating more than one race are coded as "Other." Soldiers in this sample are non-prior military service in Educational Tier 1. The sample sizes for individual criterion measures vary due to missing data.

Table 2.10. EEEM Archival Criterion Sample by MOS and Component

	Component			
MOS	Active	ARNG	USAR	Totals
11B	586	359	0	955
19K	234	73	0	313
31B	429	462	238	1,133
63B	100	125	77	307
68W	90	128	38	256
88M	120	192	65	381
AW	1,965	1,413	842	4,254
Totals	3,524	2,752	1,262	7,599

Note. The figures reported do not add up to the totals due to missing data. Soldiers in this sample are non-prior military service in Educational Tier 1.

Table 2.11. EEEM Archival Criterion Sample by MOS and Demographic Subgroup

	MOS							Totals	
Subgroup	11B	19K	31B	63B	68W	88M	AW	*n*	%
Gender									
Male	954	313	888	271	152	265	2,996	5,839	76.8
Female	0	0	244	36	104	116	1,238	1,738	22.9
Race									
White	816	272	977	247	207	270	2,981	5,770	75.9
Black	55	18	80	41	29	86	900	1,209	15.9
Other	81	21	73	17	20	23	360	595	7.8
Ethnicity									
White Non-Hispanic	731	245	888	225	193	252	2,558	5,092	67.0
Hispanic	141	37	158	38	32	38	686	1,130	14.9
AFQT Category									
I-II	332	90	373	80	196	102	1,510	2,683	35.3
IIIA	234	76	332	70	58	73	956	1,799	23.7
IIIB	354	137	415	126	2	160	1,540	2,734	36.0
IV	33	10	9	30	0	43	217	342	4.5[a]
Totals	955	313	1,133	307	256	381	4,254	7,599	

Note. The figures reported do not add up to the totals due to missing data. Soldiers indicating more than one race are coded as "Other." Soldiers in this sample are non-prior military service in Educational Tier 1. The sample sizes for individual criterion measures vary due to missing data.

[a]This number is exceeds the yearly percentage accessed into the Army because we intentionally oversampled Category IV Soldiers.

CHAPTER 3: TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM (TAPAS-95s)

Stephen Stark, O. Sasha Chernyshenko, and Fritz Drasgow
(Drasgow Consulting Group)[3]

TAPAS-95s is a new 12-dimension, 95-item personality measure, developed by Drasgow Consulting Group (DCG) under the Army's Small Business Innovation Research (SBIR) program. It was added to the Army Class predictor data collection, along with the more established AIM, as part of the EEEM project because of its potential for improving both selection and classification decisions.

Description of Measure

The TAPAS-95s builds on the foundational work of the AIM (White & Young, 1998) by incorporating features designed to promote resistance to faking and by including narrow personality constructs (i.e., facets) that are known to predict outcomes in military settings based on the most recent findings in the personality assessment and psychometric literatures.

The "s" in TAPAS-95s denotes that this is a "static" or fixed-length nonadaptive instrument, meaning that each examinee receives the same number and sequence of personality items. TAPAS-95s is designed and scored in accordance with the same psychometric models undergirding the Tailored Adaptive Personality Assessment System (TAPAS; Drasgow, Stark, & Chernyshenko, 2006; Stark, Drasgow, & Chernyshenko, 2008), which is an item response theory (IRT)-based computerized adaptive personality testing platform capable of measuring up to 22 lower-order facets of the Big Five Factor model (Goldberg, 1990), plus Physical Conditioning, which is important for military applications. Importantly, both TAPAS and TAPAS-95s utilize a multidimensional pairwise preference (MDPP) format, item response theory, and multidimensional Bayes model estimation.

A comprehensive set of 22 narrow facets of fundamental personality traits constitutes the basic building blocks of TAPAS. Rather than adhering to existing rational or theoretical nomenclature (e.g., NEO-PI or 16PF), the approach to developing the lower-order trait taxonomy was rooted in empirically examining the results of large scale factor-analytic studies, conducted using responses to a maximally diverse array of temperament indicators (e.g., adjectives, behavioral statements, or scales). Twenty-two narrow or lower-order facets were initially identified (3 – 6 facets per Big Five dimension). Within each broad Big Five domain, the lower-order facet structure was organized hierarchically. This is advantageous for applied purposes because the TAPAS system can report trait scores at any level of generality, ranging from 5 to 22 dimensions. Finally, specific to military applications, the Physical Conditioning facet was added. It was placed outside of the Big Five domain, as it is likely related to multiple factors.

Table 3.1 presents a summary of the 12 personality facets included in the current version of TAPAS-95s. The table is organized into five broad clusters representing the Big Five factors.

[3] Dr. Len White, ARI, was the Contracting Officer's Representative for the Small Business Innovation Research (SBIR) contract under which the work described in this chapter was conducted.

Within these clusters, each row presents the TAPAS facet name followed by a brief description of a typical high and/or low scorer.

Table 3.1. Description of 12 Facets Measured by TAPAS-95s

TAPAS-95s Facet	Description	Big Five Domain
Achievement	Individuals scoring high might be described as hard working, ambitious, confident, or resourceful.	Conscientiousness
Curiosity	Individuals scoring high might be characterized as inquisitive and perceptive; they read popular science/mechanics magazines and are interested in experimenting.	Openness to Experience
Non-Delinquency	Persons scoring high on this facet tend to comply with current rules, customs, norms, and expectations; they dislike change and do not challenge authority.	Conscientiousness
Dominance	High scoring individuals are domineering, take charge, and are often referred to by their peers as "natural leaders."	Extraversion
Even-Temper	Persons scoring low tend to experience a range of negative emotions including irritability, anger, hostility, and even aggression. On the other hand, persons scoring high tend to be calm, level headed, and stable.	Emotional Stability
Attention-Seeking	Individuals scoring high are constantly in search of social stimulation; they are loud, loquacious, entertaining, and even boastful.	Extraversion
Intellectual Efficiency	High scoring individuals seem to process information quickly and might be referred to by others as quick thinking, knowledgeable, astute, or intellectual.	Openness to Experience
Order	Order refers here to the ability to plan and organize tasks and activities. Persons scoring low might be referred to as disorganized, unstructured, or sloppy.	Conscientiousness
Physical Conditioning	High scoring individuals routinely participate in vigorous sports or exercise and enjoy hard physical work. On the other hand, persons scoring low are less active, and, in the extreme, might be referred to as “couch potatoes.”	Non-Big Five
Tolerance	Individuals scoring high generally enjoy cultural events and meeting and befriending people with different views. They also tend to adapt more easily to novel situations than persons scoring low.	Openness to Experience
Cooperation/Trust	High scoring individuals are trusting, cordial, cooperative, uncritical, and easy to live with, whereas those scoring low may be described as difficult, suspicious, or uncooperative.	Agreeableness
Optimism	Persons scoring high have a general emotional tone reflecting joy or happiness, whereas those scoring low have an emotional tone suggesting sadness or despair.	Emotional Stability

TAPAS-95s personality items use an MDPP format, in which items were created by pairing statements subject to similarity constraints on social desirability and/or location (extremity). A respondent's task is to choose the statement in each pair that better describes him or her. To illustrate, consider a pair of statements representing the facets of Dominance (a) and Optimism (b):

a. __ I am not one to volunteer to be group leader, but I would serve if asked.
b. __ My life has had about an equal share of ups and downs.

If we assume that a respondent's preference for a particular statement in a pair depends only on the distance from his or her standing on the trait to each statement's location on the respective trait continuum, then one can calculate the probability of preferring the statement that is closer, or more similar, to the respondent using an IRT ideal point approach. The pattern of preferences over several such items can be used to estimate a respondent's score on the various dimensions assessed by a test. Importantly, pairings can be multidimensional or unidimensional, and, in fact, a small number of unidimensional pairings is needed to identify the latent metric and obtain normative scores using the MDPP format.

The MDPP format should be more resistant to attempts at dissimulation than traditional single statement personality items. By creating tests composed of statements matched in terms of social desirability and/or location, respondents in high stakes settings should have a harder time "faking good" (Stark, Chernyshenko, & Drasgow, 2005). Another important advantage is that both static testing and adaptive testing are feasible even when only a small pool of statements is available. Because any one statement can be paired with many others, a pool of just 50 statements can produce as many as 1,225 unique pairs, if there are no constraints on repetition, location, or desirability. The TAPAS pool currently contains in excess of 1,100 statements developed by DCG to measure 22 personality facets, and approximately 200 additional statements have been made available by ARI to augment several facets, including Non-Delinquency. Consequently, even with sharp limits on repetition of statements to mitigate exposure concerns, as well as forbidding some combinations of constructs to limit faking (so-called "enemy dimensions"), the number of possible pairwise preference items available for testing is tremendous.

Development of Measure

The TAPAS-95s was developed as follows. First, Generalized Graded Unfolding Model (GGUM; Roberts, Donoghue, & Laughlin, 2000) parameters for 179 statements from 12 targeted dimensions were estimated from pre-test rating data provided by Army recruits. Next, statements similar in desirability, but representing different dimensions, were paired to form 71 multidimensional items. We also created 24 unidimensional items (two per trait) needed to identify the latent metric. Note that some pairwise preference items were pre-tested using honest and fake-good instructions and showed little score inflation. Eleven statements were used twice, so they appeared in two items. An example MDPP item from the TAPAS-95s is:

__ I hate when people are sloppy.
__ I prefer informative documentaries to other TV programs.

In this item, the first statement represents Order and the second statement represents Curiosity. Each statement has three GGUM parameters (alpha, delta, and tau) and, in an earlier study, the statements were found to be similar in social desirability. Table 3.2 presents a detailed breakdown of TAPAS-95s statements in terms of their facet designations and primary statement source (i.e., ARI or DCG statement pools).

Table 3.2. Breakdown of TAPAS-95s Statements by Facets and Source

TAPAS-95s Facet	# of Statements	Primary Statement Source
Achievement	16	ARI
Curiosity	13	DCG
Non-Delinquency	17	ARI
Dominance	17	ARI
Even-Temper	13	DCG
Attention-Seeking	14	DCG
Intellectual Efficiency	14	DCG
Order	13	DCG
Physical Conditioning	17	ARI
Tolerance	13	DCG
Cooperation/Trust	17	ARI
Optimism	15	ARI

Items selected for TAPAS-95s were randomly ordered and a paper questionnaire was created by placing five items on each page of a test booklet, preceded by an information sheet showing respondents a sample item and illustrating how to properly record their answers to the "questions" that followed. Respondents were specifically instructed to choose the statement in each pair that was "more like me" and that they must make a choice even if they found it difficult to do so. Item responses were coded dichotomously and scored using a multi-dimensional IRT method described by Stark (2002) and Stark, Chernyshenko, and Drasgow (2005).

Scoring the Measure

TAPAS-95s scoring is based on the MDPP IRT model originally proposed by Stark (2002). Rather than attempting to devise an explicit multidimensional model for pairwise preference data, we have taken a tack originally suggested by Andrich (1989, p. 197). Andrich proposed a model that assumes when person *j* encounters stimuli *s* and *t* (which, in our case, correspond to two personality statements), the person considers whether to endorse *s* and, independently, considers whether to endorse *t*. This leads to four possible outcomes: The person may wish to endorse both, neither, only *s*, or only *t*. But, when faced with a two-alternative forced choice judgment, the first two of these outcomes do not lead to a viable decision. Consequently, Andrich suggested that, in this case, the person independently reconsiders whether to endorse the two options. This process of independently considering the two stimuli continues until one and only one stimulus is endorsed. A preference judgment can then be represented by the joint outcome (Agree with *s*, Disagree with *t*) or (Disagree with *s*, Agree with *t*). Using a 1 to indicate agreement and a 0 to indicate disagreement, the outcome (1,0) indicates that statement *s* was endorsed but statement *t* was not, leading to the decision that *s* was preferred to statement *t*;

an outcome of (0,1) similarly indicates that stimulus *t* was preferred to *s*. Thus, the probability of endorsing a stimulus *s* over a stimulus *t* can be formally written as

$$P_{(s>t)_i}(\theta_{d_s},\theta_{d_t}) = \frac{P_{st}\{1,0\,|\,\theta_{d_s},\theta_{d_t}\}}{P_{st}\{1,0\,|\,\theta_{d_s},\theta_{d_t}\} + P_{st}\{0,1\,|\,\theta_{d_s},\theta_{d_t}\}},$$

where:

$P_{(s>t)_i}(\theta_{d_s},\theta_{d_t}) =$ probability of a respondent preferring statement *s* to statement *t* in item *i*,

$i =$ index for items (i.e., pairings), where $i = 1$ to I,

$d =$ index for dimensions, where $d = 1, \ldots, D$, d_s represents the dimension assessed by statement *s*, and d_t represents the dimension assessed by statement t,

$s, t =$ indices for first and second statements, respectively, in an item,

$(\theta_{d_s},\theta_{d_t}) =$ latent trait scores for the respondent on dimensions d_s and d_t respectively,

$P_{st}(1,0\,|\,\theta_{d_s},\theta_{d_t}) =$ joint probability of endorsing stimulus *s* and not endorsing stimulus *t* given latent trait scores $(\theta_{d_s},\theta_{d_t})$,

and

$P_{st}(0,1\,|\,\theta_{d_s},\theta_{d_t}) =$ joint probability of not endorsing stimulus *s* and endorsing stimulus *t* given latent trait scores $(\theta_{d_s},\theta_{d_t})$.

With the assumption that the two statements are evaluated independently, and with the usual IRT assumption that only θ_{d_s} influences responses to statements on dimension d_s and only θ_{d_t} influences responses to dimension d_t (i.e., local independence), we have

$$P_{(s>t)_i}(\theta_{d_s},\theta_{d_t}) = \frac{P_s(1\,|\,\theta_{d_s})P_t(0\,|\,\theta_{d_t})}{P_s(1\,|\,\theta_{d_s})P_t(0\,|\,\theta_{d_t}) + P_s(0\,|\,\theta_{d_s})P_t(1\,|\,\theta_{d_t})},$$

where

$P_s(1\,|\,\theta_{d_s}), P_s(0\,|\,\theta_{d_s}) =$ probability of endorsing/not endorsing stimulus *s* given the latent trait value θ_{d_s},

and

$P_t(0 \mid \theta_{d_t}), P_t(1 \mid \theta_{d_t})$ = probability of endorsing/not endorsing stimulus *t* given latent trait θ_{d_t}.

The probability of preferring a particular statement in a pair thus depends on θ_{d_s} and θ_{d_t}, as well as the model chosen to characterize the process for responding to the individual statements. Toward that end, Stark (2002) proposed using the dichotomous case of the GGUM (Roberts et al., 2000), which has been shown to fit personality data reasonably well (Chernyshenko, Stark, Drasgow, & Roberts, 2007).

Test scoring is done via Bayes modal estimation. For a vector of latent trait values, $\tilde{\theta} = (\theta_{d'=1}, \theta_{d'=2}, ..., \theta_{d'=D})$, this involves maximizing:

$$L(\tilde{u}, \tilde{\theta}) = \left\{ \prod_{i=1}^{n} \left[P_{(s>t)_i} \right]^{u_i} \left[1 - P_{(s>t)_i} \right]^{1-u_i} \right\} * f(\tilde{\theta}) \quad ,$$

where $\tilde{u}$ is a binary response pattern, $P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t})$ is the probability of preferring statement *s* to statement *t* in item *i*, and $f(\tilde{\theta})$ is a *D*-dimensional prior density distribution, which, for simplicity, is assumed to be the product of independent normals, $\prod_{d'=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\theta_{d'}^2}{2\sigma^2}}$.

Taking the natural log, for convenience, the above equation can be rewritten as:

$$\ln L(\tilde{u}, \tilde{\theta}) = \sum_{i=1}^{n} \left[(u_i) \ln P_{(s>t)_i} + (1-u_i) \ln(1 - P_{(s>t)_i}) \right] + \sum_{d'=1}^{D} \left[\ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{\theta_{d'}^2}{2\sigma^2} \right] \quad ,$$

leaving the following set of equations to be solved numerically:

$$\frac{\partial \ln L}{\partial \tilde{\theta}} = \begin{bmatrix} \frac{\partial \ln L}{\partial \theta_{d'=1}} \\ \frac{\partial \ln L}{\partial \theta_{d'=2}} \\ \vdots \\ \frac{\partial \ln L}{\partial \theta_{d'=D}} \end{bmatrix} = 0 \quad .$$

This equation can be solved numerically to obtain a vector of latent trait scores for each respondent using subroutine DFPMIN (Press, Flannery, Teukolsky, & Vetterling, 1990) in conjunction with functions that compute the posterior and its first derivatives. DFPMIN performs a *D*-dimensional *minimization* using a Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, so

the first derivatives and log likelihood values must be multiplied by –1 when maximizing. The primary advantage of this approach, over Newton-Raphson iterations, is DFPMIN does not require an analytical solution for the second derivatives.

Standard errors for TAPAS trait scores can be estimated three ways: (a) using the approximated inverse Hessian matrix that is provided by the N-dimensional minimization/ maximization routine used to compute TAPAS trait scores; (b) using a jack-knife approach, where a response pattern is scored repeatedly leaving out one item at a time and then taking the standard deviation of the resulting trait score estimates; or (c) using a new replication method, where an examinee's TAPAS trait scores are used along with parameters for the items that were administered to generate 30 new response patterns; these simulated response patterns are scored and the standard deviations of the respective trait estimates over replications are used as standard errors for the original TAPAS values. This is the method that we used here. Stark and Drasgow (2002) showed that standard errors estimated using the first method tended to be very conservative (i.e., larger than the actual empirical errors). A recent simulation by Stark and Cherynyshenko (in review) showed the second method yielded similar results, but the new replication method provided standard error estimates that were much closer to the empirical (true) standard deviations over replications.

Basic Descriptive Statistics and Psychometric Properties of Measure

Tables 3.3 and 3.4 report basic descriptive statistics for the TAPAS-95s. Table 3.3 shows means and standard deviations of 12 facet scores for the EEEM sample as well as five sub-samples created based on respondents' AFQT percentile score. Given the IRT-based scoring system, nearly all examinees' scores should lie between -3 and +3 and have a mean of zero. In practice, however, we would expect the observed means and variances of the trait score estimates to differ from these theoretical values due to differences in examinee characteristics and regression to the mean effects caused by Bayes modal estimation.

Consistent with expectations, respondents having higher AFQT percentile scores tend to be more achievement oriented, curious, even-tempered, intellectually efficient, and have, on average, greater optimism. However, the mean score differences between various samples are not particularly large, and standard deviations of TAPAS facet scores across samples are also very similar. This is promising from selection and classification standpoints. Note also that the majority of facet means are near zero indicating relatively low levels of score inflation. We do not report TAPAS scale reliabilities, as these are not particularly useful in the context of an IRT-scored measure. Standard errors of measurement generally vary across trait levels and, in this case, are reported for individual examinees along with trait scores.

Table 3.4 shows intercorrelations among the TAPAS facets as well as correlations with AFQT scores for the EEEM sample. As can be seen in the table, TAPAS facet correlations with AFQT scores are relatively low, indicating that the measure assessed constructs different from cognitive ability. Moreover, because inter-facet correlations are also relatively low, the potential for using a combination of TAPAS facets to obtain incremental validity for predicting Army-related criteria appears to be high.

Table 3.3. Descriptive Statistics for TAPAS-95s Scale Scores for the Full EEEM Sample and by AFQT Category

Scale	Items	*EEEM Sample*		AFQT Category									
				CAT I (93-100)		CAT II (65-92)		CAT IIIA (50-64)		CAT IIIB (31-49)		CAT IV (10-30)	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Achievement	16	.16	.63	.22	.69	.20	.64	.15	.62	.13	.62	.10	.53
Curiosity	13	-.09	.80	.49	.75	.04	.82	-.17	.79	-.24	.74	-.29	.72
Non-Delinquency	17	.12	.66	.17	.66	.16	.67	.13	.67	.06	.64	.11	.69
Dominance	17	-.14	.60	-.07	.61	-.12	.62	-.13	.61	-.17	.59	-.25	.56
Even-Temper	13	-.49	.76	-.22	.72	-.41	.75	-.54	.77	-.57	.77	-.65	.71
Attention-Seeking	14	-.13	.80	-.38	.74	-.16	.78	-.07	.83	-.06	.79	-.28	.72
Intellectual Efficiency	14	-.19	.65	.49	.61	.01	.64	-.29	.60	-.40	.55	-.45	.54
Order	13	-.03	.64	-.18	.64	-.03	.65	-.03	.64	-.01	.62	-.03	.62
Physical Conditioning	17	.13	.71	.04	.72	.17	.74	.11	.72	.11	.68	.15	.63
Tolerance	13	-.42	.67	-.36	.67	-.42	.68	-.41	.66	-.43	.68	-.43	.69
Cooperation/Trust	17	-.28	.86	-.39	.78	-.31	.86	-.24	.88	-.28	.87	-.24	.84
Optimism	15	-.07	.60	.16	.65	.03	.61	-.10	.58	-.18	.57	-.23	.54

Note. n = 3,381. Items = number of items comprising each final scale. CAT I Sample, n = 212. CAT II Sample, n = 1,076. CAT IIIA Sample, n = 801. CAT IIIB Sample, n = 1,114. CAT IV Sample, n = 157. Scores have a theoretical distribution of approximately -3 to +3.

Table 3.4. Intercorrelations among TAPAS-95s Scale Scores

Scale		1	2	3	4	5	6	7	8	9	10	11	12	13
1	Achievement													
2	Curiosity	**.22**												
3	Non-Delinquency	**.16**	**.10**											
4	Dominance	**.13**	**.14**	.00										
5	Even-Temper	**.05**	**.22**	**.11**	**-.06**									
6	Attention-Seeking	**-.12**	**-.12**	**-.37**	**.14**	**-.12**								
7	Intellectual Efficiency	**.16**	**.35**	.03	**.15**	**.15**	**-.08**							
8	Order	**.17**	**.05**	**.14**	**.06**	-.01	**-.08**	**.07**						
9	Physical Conditioning	**.18**	.02	**-.11**	**.05**	-.01	**.11**	.02	**.05**					
10	Tolerance	**.06**	**.20**	**.05**	**.10**	**.07**	**-.04**	**.14**	**.07**	.00				
11	Cooperation/Trust	-.01	**-.07**	**.20**	**-.13**	**.14**	**-.06**	**-.08**	.02	**-.13**	**-.03**			
12	Optimism	**.06**	**.11**	.00	**.07**	**.22**	-.03	**.19**	-.02	**.06**	**.08**	**.10**		
13	AFQT Score	**.06**	**.24**	**.06**	**.06**	**.14**	**-.07**	**.38**	**-.04**	.00	.02	**-.04**	**.18**	

Note. n = 3,381. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

CHAPTER 4: SCORING AND PSYCHOMETRIC PROPERTIES OF MEASURES

Matthew T. Allen, Yuqiu A. Cheng, and Michael J. Ingerick
(HumRRO)

This chapter begins with an overview of the scoring and psychometric properties of the criterion measures used in the EEEM analyses. A complete description of the criterion measure development can be found in Moriarty et al. (2009). The chapter concludes with the scoring and psychometric properties of the predictor measures. Note that the psychometric properties of both the criterion and predictor scores are highly similar to those reported for the full Army Class/EEEM sample reported by Allen, Cheng, Ingerick, and Caramagno (2009).

Criterion Measure Scores and Associated Psychometric Properties

Job Knowledge Tests

A single, overall score was created for each MOS-specific JKT. Obtaining this score first involved computing and analyzing standard item statistics (e.g., *p*-values, item-total correlations) to identify poorly performing items. Poorly performing items were flagged and then reviewed by the lead JKT developer to make the final determination if the item should be dropped when computing a total score. Next, a raw total score was computed by summing the total number of points Soldiers earned across the final set of items retained for each JKT. All of the multiple-choice items were worth one point. Depending on the format of the non-traditional items (e.g., multiple response, drag and drop, rank order), they were worth one or more points. To facilitate comparisons across MOS, we computed a percent correct score based on the maximum number of points that could be obtained on each MOS test. For the criterion-related validity analyses, we converted the total raw score to a standardized score (or *z*-score) by standardizing the scores *within* each MOS.

Table 4.1 shows the descriptive statistics for the raw and percent correct scores, as well as internal consistency reliability estimates for the four MOS-specific JKTs used in the EEEM analyses. Based on percent correct scores, which ranged from 55.5% (63B) to 63.9% (19K and 31B), it is evident that the tests were fairly difficult. The internal consistency reliability estimates for the JKTs were acceptable, though the 19K estimate of .63 was lower than would ordinarily be expected with this test method.

Table 4.1. Descriptive Statistics and Reliability Estimates for Job Knowledge Tests (JKTs)

MOS	*n*	Min	Max	Max Possible	*M*	*SD*	Mean Percent Correct	α
11B – Infantryman	290	46	91	118	70.57	9.75	59.8	.73
19K – Armor Crewman	228	20	54	60	38.34	5.75	63.9	.63
31B – Military Police	494	67	137	168	107.31	11.64	63.9	.72
63B – Light Wheel Vehicle Mechanic	89	39	99	122	67.65	12.04	55.5	.82

Note. Max Possible = Maximum possible score on JKT; Percent Correct = Average percent correct received on JKT [*M* / Max Possible]; α = internal consistency reliability estimate (coefficient alpha).

Rating Scales

One score was created for each dimension on the Army-wide Performance Rating Scales (AW PRS) and one overall score was created for the MOS-specific Performance Rating Scales (MOS PRS. This was done in five steps. First, the ratings were cleaned to eliminate score from raters that did not appear to be taking the task seriously. This was done by (a) checking the problem logs completed by the session proctors, (b) eliminating ratings from raters that had more than 10% missing data, (c) eliminating ratings from raters that marked more than 50% of their ratings as "Not Applicable," and (d) eliminating ratings from raters that gave the same ratings to all of the Soldiers they rated.[4] Second, average peer rating scores on each scale were computed. For example, if a Soldier was rated by three peers, for each rating scale an average score was created by computing a mean of those three scores. Third, supervisor rating scores were computed using the same procedure as what was done for the peer ratings. Fourth, peer and supervisor rating scale dimension scores were computed. This was done by taking the mean scores of all of the scales in a dimension (e.g., the three scales that describe Effort in the AW PRS), and computing an overall mean score. Finally, for each dimension, the peer and the supervisory rating scales were again averaged to create one score for each dimension.

Descriptive statistics and estimates of interrater reliability for the AW PRS dimensions and MOS PRS composite scores are shown in Appendix A (Table A.1). The interrater reliability estimates were lower than desired but consistent with our experience with the rating scales used in the Select21 concurrent validation (Ingerick, Diaz, & Putka, 2009; Knapp & Tremble, 2007). Intercorrelations among the scales are provided in Table A.2. The 11B (Infantryman) and 31B (Military Police) MOS-specific composite ratings showed generally higher correlations with the Army-wide dimensions than did those in the other MOS. The 19K (Armor Crewman) MOS-specific ratings showed the lowest correlations with the Army-wide scales.

Army Life Questionnaire

Each ALQ scale was scored differently depending on the nature of the attribute being measured. The Army Physical Fitness Test (APFT) score was unchanged. Disciplinary Incidents was recoded as a dichotomous variable, with those Soldiers self-reporting one or more incidents during IET being coded as 1 (yes) and those who did not being coded as 0 (no). The remaining four scales – Attrition Cognitions, Career Intentions, Army Fit, and Affective Commitment – were all scored with items that ranged from 1 (*strongly disagree*) to 5 (*strongly agree*). Some of the items needed to be reverse-scored. Final scores were created for these remaining scales by computing the mean of the items.

Appendix A (Table A.3) shows descriptive statistics and internal consistency reliability (coefficient alpha) estimates for the ALQ scores by MOS and for the total for-research-only criterion data EEEM sample. The reliability estimates were good (ranging from .79 to .94). Mean scores were generally similar across MOS, with the exception of the MOS Fit scale. The Military

[4] This last data screen only applied to peers and supervisors that had rated at least three Soldiers. Supervisors that rated more than 30 Soldiers were also exempted from this data screen because they were likely to have assigned the same ratings to a least three Soldiers by virtue of the number of ratings that they completed. The data from supervisors rating more than 30 Soldiers were examined closely, with information from the problem logs and the other data screens, to ensure they were not problematic.

Police (31B) results were on the higher end of the number of disciplinary incidents, but the mean number for this MOS was still quite low. Score intercorrelations for the EEEM sample are shown in Table A.4.

Six-Month Attrition

A 6-month attrition variable was computed using archival data. For purposes of this research, attrition was defined as separations due to underage enlistment, conduct, family concerns, sexual orientation, drugs/alcohol, performance, physical standards/weight, mental disorder, or violations of the Uniform Code of Military Justice. Soldiers in the dataset that had less than 6 months time-in-service were omitted from the analysis. USAR and ARNG Soldiers were also excluded because of limited availability of reliable separation data. Table 4.2 shows 6-month attrition rates for the total archival Regular Army sample and for each MOS.

Table 4.2. Attrition Rates through Six Months of Service by MOS

MOS	*N*	*N Attrit*	*% Attrit*
Total Sample	3,217	326	10.1
MOS			
11B – Infantryman	587	103	17.5
19K - Armor Crewman	233	15	6.4
31B - Military Police	427	25	5.9
63B - Light Wheel Vehicle Mechanic	92	13	14.1
68W - Health Care Specialist	89	15	16.9
88M - Motor Transport Operator	104	15	14.4
AW - Army Wide	1,685	140	8.3

Note. The statistics reported are based on Regular Army Soldiers only. *N* = number of Soldiers with 6-month attrition data. *N Attrit* = number of Soldiers who attrited through 6 months of service. *% Attrit* = percentage of Soldiers who attrited through 6 months of service [(*N Attrit*/ *N*) x 100].

IET School Performance and Completion

Data on IET school performance and completion were extracted from the ATTRS and RITMS databases. For the first variable, Graduation from AIT/OSUT, any Soldier who was discharged from Army during reception, basic training, or AIT/OSUT was coded as 0 (discharged). Any Soldier who graduated from AIT/OSUT was coded as 1 (graduated from AIT/OSUT). Any Soldier who was discharged during reception, basic training, or AIT/OSUT for nonpejorative, nonacademic reasons was excluded from the analyses. The second variable, Number of Recycles, was created by counting total number of times a Soldier was recycled to begin training again. For the third variable, Exam Grade, the average score across all exam blocks in AIT/OSUT was calculated for each Soldier. Then the standardized average score (*z*-score) was computed for each Soldier *within* MOS.

Table 4.3 shows descriptive statistics for the IET variables in the EEEM sample. The overall graduation rate was 90.2%. The lowest graduation rate was reported for 68W Soldiers because most of the sample was still in training. It is important to note that the IET data retrieved from archival

sources was not complete. For example, although there were 8,103 Soldiers in the predictor sample, we retrieved graduation data on less than 5,300 and school exam scores on less than 1,200.

Table 4.3. Descriptive Statistics for Archival IET School Performance Criteria

Graduation from AIT/OSUT	*N*	*N Grad*	*% Grad*
Total Sample	5,259	4,741	90.2
MOS			
11B – Infantryman	891	777	87.2
19K - Armor Crewman	283	283	100.0
31B - Military Police	1,100	1,052	95.6
63B - Light Wheel Vehicle Mechanic	207	188	90.8
68W - Health Care Specialist	36	16	44.4
88M - Motor Transport Operator	252	221	87.7
AW – Army Wide	2,490	2,204	88.5
Number of Recycles through AIT/OSUT	*N*	*M*	*SD*
Total Sample	7,368	.08	.30
MOS			
11B – Infantryman	936	.08	.31
19K - Armor Crewman	310	.10	.30
31B - Military Police	1,118	.02	.12
63B - Light Wheel Vehicle Mechanic	304	.08	.29
68W - Health Care Specialist	249	.22	.49
88M - Motor Transport Operator	372	.08	.29
AW – Army Wide	4,079	.08	.32

Note. *N* = number of Soldiers with data on the selected criterion. *N Grad* = number of Soldiers who completed Basic Combat Training (BCT) and graduated from AIT/OSUT. *% Grad* = percentage of Soldiers who completed BCT and graduated from AIT/OSUT [(*N Grad*/ *N*) x 100]. AIT = Advanced Individual Training; OSUT = One Station Unit Training. Most of the 68W Soldiers were still in training.

Predictor Measure Scores and Associated Psychometric Properties

Armed Services Vocational Aptitude Battery (ASVAB)

Soldiers' AFQT and ASVAB scores were extracted from Military Entrance Processing Command (MEPCOM) records and did not require any transformations or modifications. Descriptive statistics and score intercorrelations are provided in Appendix B (Tables B.1 and B.2, respectively).

Assessment of Individual Motivation (AIM)

For each AIM item tetrad, respondents provided two responses—one indicating the statement that is *most* like them and one indicating the statement that is *least* like them. A quasi-ipsative scoring method was used to generate four construct scores for each item (i.e. one score for each stem) based on whether the respondents indicated the stem is most like them, least like them, or neither. Scale scores were obtained by averaging–across items–the scores for stems measuring the same construct. A minimum of 80% of the items for any given construct must have been completed in order to obtain a score for that scale. Descriptive statistics and reliability

estimates for the AIM scales are presented in Appendix B (Table B.3). The reliability estimates were all quite acceptable (ranging from .70 to .77). The validity (i.e., lie scale) score was low, suggesting response distortion due to socially desirable responding was minimal in this sample. Table B.4 shows the intercorrelations among the AIM scale scores and the validity scale.

Rational Biodata Inventory (RBI)

RBI scores were computed by summing responses to the items applicable to each scale (reverse-scoring as required) and dividing by the number of items in the scale. Substantive scale scores were not adjusted using the response distortion scale score. Descriptive statistics and reliability estimates are shown in Appendix B (Table B.5). Most of the reliability estimates approached or exceeded .70. The substantive scales with fairly low internal consistency reliability estimates were Narcissism (.55) and Gratitude (.42). These reliability estimates, as well as the mean scores, were generally similar to what was observed with the same version of the RBI used in the Select21 concurrent validation (Knapp & Tremble, 2007), with the highest score in both samples being Self-Efficacy and the lowest score being Hostility to Authority. Scale intercorrelations are provided in Table B.6.

Predictor Situational Judgment Test (PSJT)

For each PSJT item, the respondents rated the effectiveness of four possible actions in response to a hypothetical situation. The ratings were made on a 1 (*ineffective*) to 7 (*very effective*) response scale. The PSJT was scored in the manner developed and described by Waugh and Russell (2005) and Knapp and Heffner (2009). The mean PSJT score for the total sample was 4.69 (SD = .40, n = 3,996) and the coefficient alpha reliability estimate was .86. These results are very consistent with those obtained from the Select21 and Army Class concurrent validation samples (Ingerick et al., 2009; Waugh & Russell, 2005).

Army Knowledge Assessment (AKA)

The AKA yields six dimension scores (corresponding to each of the six RIASEC dimensions). Items for each scale were averaged to create a total score for that scale. Total scores on each facet ranged from one to five. Descriptive statistics and reliability estimates for the AKA scales are shown in Table B.7. With the exception of Realistic Interests, which had a reliability estimate of .76, estimates for the remaining scales were high, ranging from .82 to .89. The scale with the highest mean score, not surprisingly for a sample of Soldiers, was Realistic Interests. AKA scale intercorrelations are shown in Table B.8.

Work Preferences Assessment (WPA)

The WPA yields six raw dimension scores (corresponding to each of the six RIASEC dimensions) and 14 facet scores (corresponding to the subfacets underlying the six RIASEC dimensions). Raw scale scores were computed by obtaining the average of the scores across the items constituting each dimension or facet. Total raw scale scores range from one to five. Alternative algorithms for scoring the WPA are available, including algorithms that factor in environment or job-side data on the kinds of work activities and settings supported by the Army in general or a specific job. Only the raw scale scores were used in the current research because (a) past research has shown that alternative scoring algorithms produce comparable criterion-

related validity estimates and (b) the empirically-keyed scoring algorithms were developed under a concurrent validation design and using criterion data that were collected in-unit and not at the end of Soldiers' IET.

Descriptive statistics and reliability estimates for both the dimension and facet scores are shown in Table B.9 in Appendix B. Most reliability estimates were relatively high (mid-.70s to .90). Several of the facet scores were a bit lower, with Clear Procedures (a facet of Conventional Interests) being the score with the lowest estimated reliability (.65). The WPA score intercorrelations are shown in Table B.10.

CHAPTER 5: ANALYSIS AND FINDINGS

Matthew T. Allen, Yuqiu A. Cheng, Dan J. Putka (HumRRO),
Arwen Hunter, and Len White (ARI)

Introduction

In this chapter, we first summarize the results of our analyses to determine which of the experimental predictor measures represented the "best bets" for enhancing new Soldier recruitment and selection. Next, we review the results and findings of analyses to construct candidate performance screens based on selected "best bet" predictors. Finally, the chapter concludes with a discussion of the construction and evaluation of a performance screen for use in an IOT&E.

Selection of "Best Bet" Experimental Predictor Measures

Approach

Four factors were considered in evaluating which experimental predictor measures represented the "best bets" for enhancing new Soldier selection: (a) incremental validity (over the AFQT), (b) subgroup differences, (c) susceptibility to faking and coaching effects, and (d) administration time. The first factor, incremental validity, was estimated by computing the increment in multiple correlation (ΔR) over the AFQT when each of the experimental predictor measures was added to a regression model predicting a valued outcome (e.g., performance, retention). All other factors being equal, the greater the incremental validity from adding the new predictor measure(s), the greater the potential to enhance new Soldier screening. Estimating the incremental validity of the experimental predictor measures involved fitting a series of hierarchical regression models. In the first step, each criterion was regressed on Soldiers' AFQT scores. In the second step, the criterion was regressed on Soldiers' scale scores from the experimental predictor measure in addition to their AFQT scores.

Ordinary least squares (OLS) regression was used to estimate incremental validity for the continuously-scaled criteria (e.g., job knowledge, effort). For the dichotomously-scored criteria (i.e., attrition and disciplinary incidents), the same hierarchical, multi-step approach was followed, but using logistic regression. Since there is no effect size estimate in logistic regression directly equivalent to ΔR, the capacity of the experimental measure to add incremental validity beyond AFQT was assessed using three statistics. The first was Nagelkerke's R, which provides a pseudo estimate of R. The second statistic was the standardized mean difference (or Cohen's d) in the predicted probabilities between Soldiers who experience an event (i.e., attriting or having a disciplinary incident) versus those who did not, using the AFQT and experimental measures as predictors. The larger the d value the greater the difference in predicted probabilities between those who did and did not experience an event (e.g., attrited). The better experimental predictor measure(s) will evidence a higher d value than that obtained using the AFQT only. The third statistic was a point-biserial correlation between Soldiers' predicted probability of a negative event (e.g., attriting or having a disciplinary incidence), based on the AFQT and the experimental predictor measures, and their actual behavior. Similar to the standardized mean difference, the

better experimental predictor measure(s) will yield higher point-biserial correlations than those obtained using the AFQT only.

To ensure the comprehensiveness of the incremental validity analyses, the analyses were conducted on 14 of the available criterion measures, representing two types of criteria: (a) performance-related and (b) retention-related. The *performance-related* criteria were chosen based on work described in Campbell, Hanson, and Oppler (2001) on the latent factor structure of first-term Army job performance (see also Campbell, McHenry, & Wise, 1990). At least two criterion measures were chosen to represent each of the Campbell et al. performance dimensions, excluding General Soldiering Proficiency which was assessed with only one measure:

1. *Core Technical Proficiency* – Core Technical Proficiency represents the extent to which Soldiers perform the tasks that are essential to their MOS. This dimension was assessed using (a) the MOS-specific JKT and (b) the MOS-specific performance rating composite.

2. *Effort and Leadership* – This dimension reflects the extent to which the Soldier perseveres in the face of adversity and supports other Soldiers. Effort and Leadership was measured using three Army-wide performance rating scales: (a) Effort, (b) Support for Peers, and (c) Peer Leadership.

3. *Maintaining Personal Discipline* – Maintaining Personal Discipline reflects the extent to which Soldiers demonstrate commitment and discipline. This dimension was assessed using (a) the Personal Discipline Army-wide performance rating scale and (b) the occurrence of a disciplinary incidents during IET (yes/no), as self-reported on the ALQ.

4. *Physical Fitness and Military Bearing* – This dimension represents the extent to which a Soldier maintains an appropriate Army appearance and good physical condition. It was measured using (a) the Physical Fitness and Bearing Army-wide performance rating scale and (b) the Soldiers' most recent APFT score, as self-reported on the ALQ.

The *retention* criteria included (a) Soldier attrition through the first 6 months of service and (b) a series of attitudinal retention criteria. The attitudinal retention criteria were chosen based on previous research showing which Soldier attitudes were most predictive of attrition and first-term re-enlistment behavior (e.g., Strickland, 2005) and were measured by scales administered in the ALQ. The attitudes selected were as follows (see Table 2.3 for more information):

1. Affective Commitment
2. Army Fit
3. Career Intentions
4. Attrition Cognitions

The second factor considered when evaluating the "best bets" predictors, subgroup differences, was evaluated by computing the standardized mean differences between targeted demographic subgroups in the scale scores on the experimental predictor measures. The demographic subgroups targeted for our analyses were (a) gender (female-male), (b) race (Black-White), and (c) ethnicity (Hispanic-White, Non-Hispanic). Standardized mean differences were computed using a variant of Cohen's *d* statistic:

$$d = (M_{COMPARISON} - M_{REFERENT})/SD_{REFERENT}.$$

where

M = Group Mean
SD = Group Standard Deviation.

For the purpose of this analysis, the comparison groups were Female, Black, and Hispanic, while the referent groups were Male, White, and Non-Hispanic, respectively.

Empirical data were not collected during this research to evaluate the experimental predictor measures on the final two factors – susceptibility to faking/coaching effects and administration time. Accordingly, the experimental predictor measures' susceptibility to faking and coaching effects were evaluated using a combination of rational judgment and findings from previous research. The time allotted for completing each measure was used to estimate administration times. Only measures that showed promise on the first two factors were examined on these final two factors. The results and findings from our evaluation are described next.

Findings

Appendix C shows the uncorrected bivariate scale-level correlations between selected predictor and criterion measures.

Incremental Validity

Results of the incremental validity analyses are summarized in Tables 5.1 to 5.3. These analyses yielded consistent findings regarding which experimental predictor measures represent the best bet predictors for new Soldier selection beyond AFQT. In regards to performance-related criteria (Table 5.1), three measures – RBI (ΔR = .043-.228), TAPAS (ΔR = .030-.194), and AIM (ΔR = .016-.171) – emerged as the strongest incremental predictors among the experimental predictor measures. All three measures evidenced ΔR statistics that were consistently higher than the other four experimental measures. The AKA and PSJT (average ΔR = .015 and .006, respectively) tended to predict the least amount of incremental variance in the performance criteria beyond AFQT.

The same experimental measures that emerged as "best bet" predictors for the performance-related criteria also emerged as strong incremental predictors of retention-related criteria (Table 5.2). RBI (ΔR = .248-.386), AIM (ΔR = .214-.341), and TAPAS (ΔR = .179-.289) consistently emerged as the measures exhibiting the greatest incremental validity beyond AFQT. The WPA (ΔR = .197-.317) also predicted a substantial amount of incremental variance in the retention-related criteria when computed at the facet, rather than dimension, level. Given that the retention-related criteria are more attitudinal rather than performance-oriented, it should come as no surprise that the more cognitively-oriented experimental measures – AO (ΔR = .004-.030) and PSJT (ΔR = .002-.040) – did not predict retention-related attitudes beyond AFQT.

As Table 5.3 shows, the same four predictors of retention-related criteria – RBI, TAPAS, AIM,and WPA – were also the four experimental measures most likely to incrementally predict 6-month attrition. Among these, the RBI and the TAPAS evidenced the greatest incremental validity for predicting 6-month attrition beyond AFQT. Similarly, the TAPAS, followed closely by the RBI, demonstrated the greatest incremental validity for predicting disciplinary incidents.

Table 5.1. Incremental Validity Estimates for Experimental Predictors over the AFQT for Predicting Performance-Related Criteria

Criterion/Predictor	*n*	AFQT Only	AFQT + Predictor	Δ*R*
MOS-Specific Job Knowledge Test				
AO [1]	972	**.476**	**.487**	**.011**
AIM [6]	355	**.476**	**.492**	.016
TAPAS [12]	504	**.476**	**.506**	.030
PSJT [1]	695	**.476**	**.485**	**.010**
RBI [14]	796	**.476**	**.518**	**.043**
AKA [6]	1,058	**.476**	**.488**	**.012**
WPA Dimensions [6]	1,050	**.476**	**.501**	**.026**
WPA Facets [14]	1,050	**.476**	**.511**	**.036**
MOS-Specific Performance Ratings Composite				
AO [1]	1,028	**.136**	**.179**	**.043**
AIM [6]	374	**.136**	**.198**	.062
TAPAS [12]	536	**.136**	**.237**	.101
PSJT [1]	729	**.136**	**.145**	.009
RBI [14]	824	**.136**	**.219**	**.083**
AKA [6]	1,108	**.136**	**.161**	.026
WPA Dimensions [6]	1,103	**.136**	**.176**	**.040**
WPA Facets [14]	1,103	**.136**	**.190**	.054
Effort Ratings Composite (Army-Wide)				
AO [1]	1,049	**.184**	**.221**	**.037**
AIM [6]	377	**.184**	**.231**	.047
TAPAS [12]	538	**.184**	**.266**	.082
PSJT [1]	752	**.184**	**.193**	.009
RBI [14]	848	**.184**	**.255**	**.071**
AKA [6]	1,134	**.184**	**.192**	.008
WPA Dimensions [6]	1,128	**.184**	**.197**	.013
WPA Facets [14]	1,128	**.184**	**.213**	.028
Physical Fitness and Bearing Ratings Composite (Army-Wide)				
AO [1]	1,049	**.080**	**.125**	**.045**
AIM [6]	377	.080	**.251**	**.171**
TAPAS [12]	538	.080	**.274**	**.194**
PSJT [1]	752	**.080**	**.081**	.001
RBI [14]	848	**.080**	**.309**	**.228**
AKA [6]	1,134	**.080**	**.098**	.018
WPA Dimensions [6]	1,128	**.080**	**.139**	**.059**
WPA Facets [14]	1,128	**.080**	**.161**	.081

Table 5.1. Incremental Validity Estimates for Experimental Predictors over the AFQT for Predicting Performance-Related Criteria (cont'd)

Criterion/Predictor	*n*	AFQT Only	AFQT + Predictor	Δ*R*
Support for Peers Ratings Composite (Army-Wide)				
AO [1]	1,049	**.145**	**.163**	**.018**
AIM [6]	377	**.145**	**.246**	**.101**
TAPAS [12]	538	**.145**	**.238**	.093
PSJT [1]	752	**.145**	**.147**	.002
RBI [14]	848	**.145**	**.239**	**.094**
AKA [6]	1,124	**.145**	**.167**	.022
WPA Dimensions [6]	1,128	**.145**	**.165**	.020
WPA Facets [14]	1,128	**.145**	**.191**	.045
Leadership Ratings Composite (Army-Wide)				
AO [1]	1,048	**.136**	**.177**	**.041**
AIM [6]	377	**.136**	**.265**	**.129**
TAPAS [12]	537	**.136**	**.273**	**.136**
PSJT [1]	752	**.136**	**.141**	.005
RBI [14]	848	**.136**	**.263**	**.127**
AKA [6]	1,133	**.136**	**.139**	.003
WPA Dimensions [6]	1,128	**.136**	**.158**	.022
WPA Facets [14]	1,128	**.136**	**.171**	.035
Discipline Ratings Composite (Army-Wide)				
AO [1]	1,049	**.172**	**.188**	**.015**
AIM [6]	377	**.172**	**.295**	**.123**
TAPAS [12]	538	**.172**	**.332**	**.160**
PSJT [1]	752	**.172**	**.183**	.010
RBI [14]	848	**.172**	**.270**	**.098**
AKA [6]	1,134	**.172**	**.180**	.007
WPA Dimensions [6]	1,128	**.172**	**.193**	.021
WPA Facets [14]	1,128	**.172**	**.218**	.046
Army Physical Fitness Test (APFT)				
AO [1]	1,002	.035	.037	.002
AIM [6]	356	.035	**.299**	**.264**
TAPAS [12]	521	.035	**.298**	**.263**
PSJT [1]	734	.035	.040	.004
RBI [14]	818	.035	**.379**	**.344**
AKA [6]	1,091	.035	.060	.024
WPA Dimensions [6]	1,092	.035	.071	.036
WPA Facets [14]	1,092	.035	**.168**	**.133**

Note. AFQT = Armed Forces Qualification Test. *AFQT Only* = Correlation between the AFQT and the criterion. *AFQT + Predictor* = Multiple correlation (*R*) between the AFQT and the selected predictor measure with the criterion. Δ*R* = Increment in *R* over the AFQT from adding the selected predictor measure to the regression model (AFQT + Predictor – AFQT Only). Estimates in bold are statistically significant, $p < .05$ (two-tailed). The numbers in brackets after the title of the predictor measure indicate the number of scale scores that the measure contributed to the regression model. The WPA consists of six dimensions and 14 facets embedded within those dimensions.

Table 5.2. Incremental Validity Estimates for Experimental Predictors over the AFQT for Predicting Retention-Related Criteria

Criterion/Predictor	*n*	AFQT Only	AFQT + Predictor	Δ*R*
Affective Commitment				
AO [1]	1,003	.062	.065	.004
AIM [6]	356	.062	**.303**	**.241**
TAPAS [12]	522	.062	**.246**	**.184**
PSJT [1]	733	.062	.064	.002
RBI [14]	820	.062	**.431**	**.369**
AKA [6]	1,092	**.062**	**.188**	**.126**
WPA Dimensions [6]	1,091	**.062**	**.261**	**.200**
WPA Facets [14]	1,091	**.062**	**.286**	**.224**
Needs-Supplies Army Fit				
AO [1]	1,001	.008	.037	.030
AIM [6]	352	.008	**.349**	**.341**
TAPAS [12]	516	.008	**.281**	**.273**
PSJT [1]	736	.008	.048	.040
RBI [14]	820	.008	**.393**	**.386**
AKA [6]	1,089	.008	**.159**	**.151**
WPA Dimensions [6]	1,091	.008	**.265**	**.257**
WPA Facets [14]	1,091	.008	**.325**	**.317**
Career Intention				
AO [1]	1,003	.014	.037	.024
AIM [6]	354	.014	**.262**	**.249**
TAPAS [12]	518	.014	**.302**	**.289**
PSJT [1]	736	.014	.021	.007
RBI [14]	819	.014	**.317**	**.304**
AKA [6]	1,091	.014	**.141**	**.128**
WPA Dimensions [6]	1,090	.014	**.222**	**.209**
WPA Facets [14]	1,090	.014	**.263**	**.250**
Attrition Cognition				
AO [1]	1,000	**.069**	.076	.007
AIM [6]	355	.069	**.283**	**.214**
TAPAS [12]	515	.069	**.248**	**.179**
PSJT [1]	731	.069	.077	.009
RBI [14]	818	**.069**	**.317**	**.248**
AKA [6]	1,087	**.069**	**.169**	**.101**
WPA Dimensions [6]	1,087	**.069**	**.187**	**.118**
WPA Facets [14]	1,087	**.069**	**.265**	**.197**

Note. AFQT = Armed Forces Qualification Test. *AFQT Only* = Correlation between the AFQT and the criterion. *AFQT + Predictor* = Multiple correlation (*R*) between the AFQT and the selected predictor measure with the criterion. Δ*R* = Increment in *R* over the AFQT from adding the selected predictor measure to the regression model (AFQT + Predictor – AFQT Only. Estimates in bold are statistically significant, $p < .05$ (two-tailed). The numbers in brackets after the title of the predictor measure indicate the number of scale scores that the measure contributed to the regression model. The WPA consists of six dimensions and 14 facets embedded within those dimensions.

Table 5.3. Incremental Validity, Cohen's d, and Point-Biserial Estimates for Experimental Predictors over the AFQT for Predicting Dichotomous Criteria

		Incremental Validity			*Cohen's d*			*Point-Biserial (r_{pb})*		
Predictor	*n*	AFQT Only	AFQT + Predictor	Δ *R*	AFQT Only	AFQT + Predictor	Δ *d*	AFQT Only	AFQT + Predictor	Δ r_{pb}
6-Month Attrition										
AO [1]	2,960	.030	**.097**	**.066**	.068	.250	.182	.021	.074	.053
AIM [6]	1,633	.052	**.215**	**.162**	.111	.518	.407	.037	.164	.127
TAPAS [12]	1,689	.049	**.243**	**.195**	.107	.624	.517	.034	.183	.150
PSJT [1]	1,252	.041	.093	.052	.103	.247	.145	.026	.061	.035
RBI [14]	2,478	.050	**.243**	**.193**	.110	.718	.607	.035	.201	.166
AKA [6]	2,975	.029	.096	.067	.065	.238	.172	.020	.070	.051
WPA Dimensions [6]	2,955	.029	**.124**	**.096**	.065	.309	.245	.020	.091	.071
WPA Facets [14]	2,953	.029	**.162**	**.133**	.065	.439	.375	.020	.125	.106
Disciplinary Incidents										
AO [1]	1,011	**.098**	**.135**	.037	.170	.260	.090	.079	.115	.036
AIM [6]	546	**.153**	.229	.076	.262	.444	.182	.123	.197	.074
TAPAS [12]	712	.085	**.221**	**.137**	.145	.443	.298	.068	.189	.121
PSJT [1]	981	**.097**	.117	.020	.164	.219	.054	.079	.100	.020
RBI [14]	1,191	**.133**	**.248**	**.115**	.228	.457	.229	.108	.207	.098
AKA [6]	1,531	**.111**	.123	.011	.192	.217	.024	.091	.102	.011
WPA Dimensions [6]	1,526	**.106**	.131	.025	.183	.234	.051	.087	.109	.022
WPA Facets [14]	1,525	**.106**	.160	.053	.184	.296	.112	.087	.134	.048

Note. Δ*R* = Increment in Nagelkerke's *R* over AFQT. Cohen's *d* = Standardized mean difference in the predicted probabilities between Soldiers who attrit/fail and those who persist/pass. r_{pb} = Point-biserial correlation between Soldiers' predicted probability of attriting/failing with their actual attrition/failure behavior. Estimates in bold are statistically significant, $p < .05$ (two-tailed). The numbers in brackets after the title of the predictor measure indicate the number of scale scores that the measure contributed to the regression model. The WPA consists of six dimensions and 14 facets embedded within those dimensions.

Subgroup Differences

In addition to examining incremental validity, we also examined subgroup differences in the scale scores by predictor measure. Table D.1 (Appendix D) reports the subgroup differences for all seven experimental predictors and AFQT. In summarizing the main findings from these analyses, we focus particular attention on the four measures demonstrating the greatest incremental validity (RBI, TAPAS, AIM, and WPA). Among these four measures, the forced-choice temperament measures (i.e., TAPAS, AIM) evidenced the smallest subgroup differences on average. Only two of the TAPAS scales exhibited a standardized difference greater than .30 (Tolerance and Non-Delinquency). On both positively-valenced scales, the minority group scored about .33 standard deviations higher on average than the majority group. The AIM also generally exhibited small subgroup differences, with absolute standardized mean differences ranging from .03 to .32 (average absolute d = .15). Same as with the TAPAS, the minority group scored higher on average than the majority group on the two scales demonstrating the largest subgroup differences (Dependability and Work Orientation).

The RBI evidenced larger subgroup differences, on average, than the AIM and TAPAS, with absolute standardized mean differences ranging from .01 to .75 (average absolute d = .19). Among the scales exhibiting the largest subgroup differences, the direction of the score differences varied. On several scales, the differences favored minority group members (e.g., Black Soldiers had higher Achievement scores than White Soldiers, d = .40), while on others they favored the majority group (e.g., Black Soldiers had higher Narcissism scores than White Soldiers, d = .46). Among all the experimental predictor measures, the RBI Fitness Motivation scale demonstrated the largest subgroup difference, with females scoring nearly three-fourths of a standard deviation lower than males (d = -.75).

The WPA exhibited a comparable pattern of subgroup differences to the RBI, at both the dimension and facet level, with absolute standardized mean differences ranging from .00 to .76 (average absolute d = .30 for dimensions, average absolute d = .26 for facets). Where there were several sizeable gender and race differences, those differences did not consistently favor one subgroup over another – some scales had larger mean differences that favored the minority group, while others favored the majority group. For example, Black Soldiers scored higher, on average, than White Soldiers on the Conventional Interests scale (d = .58), whereas White Soldiers scored higher than Black Soldiers on the Realistic Interests scale (d = -.44). Because the WPA measures attributes that are less valenced than several of the other predictor measures (i.e., a high or low score is not necessarily considered “good” or “bad”), these differences raise fewer practical concerns than differences exhibited by those other measures.

In sum, among the four experimental predictor measures showing the greatest potential for enhancing new Soldier recruitment and selection, the TAPAS and the AIM emerged as the measures evidencing the fewest subgroup differences.

Conclusions and Recommendations on "Best Bet" Experimental Predictor Measures

Table 5.4 summarizes how the experimental predictor measures compare on the four evaluation factors. On the basis of this information, several conclusions can be drawn:

1. Three measures consistently emerged as "best bets" for predicting Soldier performance and retention-related criteria. Those three measures were the RBI, TAPAS, and AIM. A fourth measure, the WPA, also showed promise.

2. Of these measures, the TAPAS and AIM evidenced the smallest subgroup differences, on average. The RBI exhibited the highest subgroup differences, followed by the WPA. However, in most cases, these subgroup differences were such that minority group members scored higher, on average, on the predictor measure than majority group members.

3. Among the three most promising measures (TAPAS, AIM, and RBI), the TAPAS potentially represents the measure least susceptible to faking or coaching effects in an operational, high stakes setting. The RBI is arguably the measure most susceptible to faking and coaching effects because of its self-report nature, although its inclusion and possible use of a validity scale could at least partially offset these effects. As mentioned, we were not able to examine comprehensively these issues empirically in the present research and with the current sample. Accordingly, this conclusion is based on previous research with the RBI (Kilcullen et al., 2005) and with rationally-derived biodata measures in general (Graham, McDaniel, Douglas, & Snell, 2002; Harold, McFarland, & Weekley, 2006; McFarland & Ryan, 2000). An advantage of the AIM is that it uses a forced-choice response format designed to be resistant to faking. Although initial research on the AIM's susceptibility to score inflation was promising, subsequent research showed that its criterion-related validity initially dropped under operational conditions (Knapp, Heggestad, & Young, 2004; White et al., 2008). The problem was addressed by constructing a rational-empirical keyed scoring system tailored exclusively to the prediction of first-term attrition. Following introduction of the new system, the AIM was used successfully in operational conditions (White et al., 2008). The TAPAS represents a potential enhancement over AIM, and other temperament measures, because the statements used for each item are carefully matched on both social desirability and trait location to reduce fakability and coachability. In addition, because any one statement can be paired with any other, a 15-dimension TAPAS would contain over 100,000 possible items and item exposure would be further limited by use of adaptive testing and constraints on repetition of statements. All of these factors help to limit faking and make TAPAS more suitable for implementation in any large scale, "high stakes" testing program, like preenlistment screening for the military, where faking and coaching is a concern. The adaptive TAPAS, relative to static measures, can also reduce the testing time required by tailoring the questions to the individual examinee.

4. The required administration time for the three most promising "best bet" measures (RBI, AIM, and TAPAS) was identical (30 minutes).

In sum, the TAPAS appears to represent the "best bet" predictor measure for enhancing new Soldier selection in an operational setting. It exhibited high incremental validity, few subgroup differences, has the potential to be less susceptible to faking or coaching effects than other candidate measures (e.g., AIM, RBI), and can be administered in a reasonable timeframe.

Table 5.4. Comparison of Army Experimental Predictor Measures on Important Factors

Predictor	Incremental Validity	Subgroup Differences	Response Distortion Potential	Administration Time (in minutes)
1. Rational Biodata Inventory (RBI)	HIGH	MED	HIGH	30
2. Predictor Situational Judgment Test (PSJT)	LOW	MED	LOW	30
3. Work Preferences Assessment (WPA)	MED	MED	MED	20
4. Army Knowledge Assessment (AKA)	LOW	LOW	LOW	10
5. Tailored Adaptive Personality Assessment System (TAPAS)	HIGH	LOW	LOW	30
6. Assessment of Individual Motivation (AIM)	HIGH	LOW	HIGH/LOW[a]	30

Note. AO is omitted because it is already administered as part of the ASVAB.
[a]AIM has high response distortion potential when using the original scoring system and low potential when using rational-empirical key adopted for the TTAS program.

Initial Development and Evaluation of Candidate Performance Screens

Approach

Developing Candidate Performance Screens

Developing a performance screen (or screens) that combines information on one or more of the experimental predictor measures for potential use in new Soldier recruitment and classification required making several design choices. They were (a) which predictor measure(s) (or scales within measures) to include in the screen(s), (b) which criterion measures to use when developing the screen(s), (c) how many screens to develop, and (d) what is the process or procedures to be used to determine which scales to include in the screen(s).

Based on the findings reported in the preceding section, the TAPAS emerged as the "best bet" candidate for a new screening instrument. Accordingly, the TAPAS was selected as the measure to be used in constructing candidate performance screens.

The second issue was what criteria to use in constructing performance screens based on the TAPAS, as this has implications for which scales to include. Previous Army research has analyzed criteria individually (e.g., Ingerick et al., 2009) or formed criterion composites (e.g., Knapp & Tremble, 2007). The difficulty with the first approach is the number of criteria used in this project would quickly make analyses impractical for interpretation. With the second approach, forming criterion composites would combine the errors of the different measurement methods, which can further contaminate the results. As a compromise to these two approaches,

we focused our analysis on maximizing (or minimizing) five criteria valued by the Army and that collectively provide reasonably comprehensive coverage of the criterion space: (a) the PRS scale measuring Effort, (b) self-reported APFT score, (c) 6-month attrition, (d) self-reported disciplinary incidents, and (e) the MOS-specific JKT.

The third issue was the number of performance screens to develop. One option was to develop a single composite of TAPAS scales that maximizes scores on all of the selected criteria. The alternative was to develop multiple performance screens, each targeted toward a single criterion. The advantage to the second approach is its flexibility downstream. For example, one or more of the developed screens could be applied, as needed, depending on Army recruitment priorities. A "final" performance screen could then be created by requiring Soldiers to receive a score above a certain level on a percentage of the composites (e.g., two out of five). This flexibility in scoring is the reason we developed separate composites for each of the five criteria of interest.

The fourth and final issue was how to select the scales. Two different approaches were used to construct candidate performance screens for each of the five targeted criteria. The first approach combined theory with empirical results to derive the screens. This approach consisted of the following three steps:

1. The TAPAS scales with the closest theoretical association to the five criteria were identified by project researchers. For example, it was proposed that the Achievement and Physical Conditioning scales would be most closely related Soldiers' APFT scores. Bivariate correlations between the TAPAS scales and the criteria were used to check whether the proposed direction of the relationship was supported by the data.

2. Forward stepwise regression was used to determine whether one or more of the remaining TAPAS scales added incremental variance to the existing model. For continuous criteria, OLS regression was used to regress each criterion on AFQT in the first step, then the theoretical TAPAS scales in the second step. In the third step, the remaining TAPAS scales were entered using the forward selection procedure.[5] The same approach was used for the composites developed for the dichotomous criteria, but using a logistic regression framework instead of OLS. Any scales that added a significant amount of explanatory variance were added to the composite.

3. Composite scores were computed by adding all of the selected scales that were positively correlated with the criterion and subtracting all of the scales that were negatively correlated with the criterion.

[5] In the forward selection procedure, a scale is only added to the model if the *F*-value for a given scale is significant at $p < .05$. The TAPAS scale with the strongest partial correlation with the criterion is added until the *F*-value probability for the remaining TAPAS scales exceeded .05.

The second approach used to derive TAPAS-based performance screens was purely empirical and consisted of the following steps:

1. The initial TAPAS scales included in the model were determined by using the "best subsets" regression procedure. In this analysis, all possible variations of a set of predictors (in this case, the 12 TAPAS scales and AFQT) were computed and evaluated using some statistic (in this case, Mallow's C_p[6]). A scale was included in a composite if it was present in five or more of the "best" 10 models.

2. As with the combined theoretical and empirical approach described above, forward stepwise regression was used to determine whether one or more of the remaining TAPAS scales added incremental variance to the existing model.

3. After finalizing which scales were to be included in the screens, composite scores were computed by adding and subtracting scores on the TAPAS scales, depending on their relationship with the criterion. However, instead of using bivariate correlations (as was the case with the combined approach), standardized betas from the OLS regression analyses were used to determine whether a scale was weighted positively or negatively when forming the composite.

Evaluating Candidate Performance Screens

The combined theoretical-empirical and purely empirical approaches were evaluated using results from two sets of analyses. The first was the OLS regression results. As was the case when picking a "best bet" predictor measure, the composites that provide the most incremental validity beyond AFQT indicate they are more likely to predict enlisted Soldier performance and retention in an initial operational test and evaluation (IOT&E).

The second set of analyses used to evaluate the two types of composites was the "split-group" analyses. To use non-cognitive measures to meet mission requirements, either to expand ("select-in") or refine ("select-out") the recruiting pool, the EEEM composites have to be evaluated in the context of the AFQT categories. The basis for the group comparison was made using Cohen's *d* for continuous criteria, and a Relative Risk (RR) ratio (*p* [incident for comparison group]/ *p* [incident for referent group]) [7] for the dichotomous criteria. A final decision was then made regarding which composites should be used to maximize (or minimize) each of the five criteria.

[6] Mallows' C_p is a diagnostic statistic commonly used for regression model evaluation. While it is meant to be used as a tool for further model examination (Mallows, 1973), in general lower scores indicate a better fitting, less biased model (Zuccaro, 1992). In this analysis, all possible regression models for a particular criterion were sorted by Mallows' C_p.

[7] *p* = probability, which in RR is equivalent to the percentage of Soldiers with an incident for the referent (Category IIIA) and comparison (Category IIIB) groups.

Findings

Development of TAPAS-Based Performance Screens

A total of 10 TAPAS-based performance screens were developed, resulting in two screens for each of the five targeted criteria (one screen derived from the empirically-based approach and a second based on a combined theoretical-empirical approach). The empirically-based TAPAS screens were derived from a series of best subsets regression analyses. The "best" 10 models, as measured by Mallow's C_p, were outputted for each of the five criteria. Table 5.5 provides a summary of the results of the best subsets regression analyses. This table shows (a) the total number of "Top 10" models each scale belonged to and (b) which criterion measure that scale consistently predicted (i.e., is found in five or more of the Top 10 models). The results show that some scales, such as Physical Conditioning and Attention-Seeking, emerged as significant predictors across multiple criteria, whereas others, such as Tolerance and Coooprationn/Trust, did not emerge as predictors for any criteria.

The next step in developing empirically-based TAPAS screens was to determine whether any of the holdover TAPAS scales that were not included in the initial screen based on the best subsets analysis could explain significant incremental variance in one of the targeted criteria. As Table 5.6 shows, no TAPAS scales significantly incremented the prediction of the five targeted criteria. The sign on the resulting beta weights for each individual scale was used to determine how the scale would be weighted in the screen. For example, for the Physical Fitness screen, Achievement, Order, and Physical Conditioning were weighted positively, while Non-delinquency and attention seeking received a negative weight. Table 5.7 summarizes the composition of the final set of empirically-derived TAPAS screens.

Table 5.5. Summary of Best Subsets Analysis Results

TAPAS Scale	Number of Models	% of Models	Criterion Dimension Consistently Predicted				
			JKT	Effort	APFT	Disc	Attrit
Achievement	27	54%	X	X	X		
Curiosity	6	12%					
Non-Delinquency	12	24%			X		
Dominance	9	18%					X
Even-Temper	30	60%	X			X	X
Attention-Seeking	34	68%	X	X	X	X	
Intellectual Efficiency	20	40%		X		X	
Order	18	36%			X	X	
Physical Conditioning	32	64%	X	X	X		X
Tolerance	5	10%					
Cooperation/Trust	3	6%					
Optimism	12	24%					X

Note. Number of Models = the total number of "Top 10" models in which the scale was included, % of models = percent (out of 50) of "Top 10" models in which the scale was included. JKT = Job Knowledge Test, Effort = AW PRS measure of overall Effort, APFT = Army Physical Fitness Test, Disc = Disciplinary incidents during IET, Attrit = 6-month attrition.

Table 5.6. Incremental Validity Estimates of Empirically-Derived "Best Bet" TAPAS Scales over the AFQT for Predicting Selected Criteria

Criterion	*n*	Step 1 AFQT Only	Step 2 Selected TAPAS	Step 3 Remaining TAPAS	Δ*R* Over AFQT
MOS-Specific JKT	504	**.476**	**.493**	Done	**.017**
Army-Wide Effort PRS	538	**.184**	**.247**	Done	**.063**
Army Physical Fitness Test Score (APFT)	521	.035	**.281**	Done	**.246**
Disciplinary Incidents	522	.036	**.250**	Done	**.214**
6-Month Attrition	1,689	.049	**.236**	Cone	**.188**

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. *Step 1 AFQT Only* = Correlation between the AFQT and the criterion. *Step 2 Selected TAPAS* = Multiple correlation (*R*) between the AFQT + Selected TAPAS scales with the criterion. *Step 3 Remaining TAPAS = R* between AFQT + Selected TAPAS scales + scales included with the forward selection method. Δ*R Over AFQT* = Increment in *R* over the AFQT from adding the selected TAPAS scales to the regression model. Δ*R Over Selected TAPAS* = Increment in *R* over AFQT and selected TAPAS scales from adding additional scales with forward selection. The *R*'s for Disciplinary Incidents and 6-Month Attrition were estimated using Nagelkerke's method in logistic regression. Estimates in bold are statistically significant, $p < .05$ (two-tailed).

One issue with using a purely empirically-driven approach for deriving the TAPAS-based performance screens was some unexpected, and at times counterintuitive, results. For example, there were a number of TAPAS scales identified for inclusion in the empirically-derived screens that evidenced small standardized beta weights and where there was little theoretical explanation for their inclusion. For example, the Attention-Seeking TAPAS scale emerged as a significant predictor for the MOS-specific JKTs. However, there is little theoretical justification for relating physical condition to Soldier's technical job knowledge; indeed, the bivariate correlation between the two variables was near zero (see Table 5.8). By contrast, other TAPAS scales were notably absent from selected screens. For example, one would expect the Non-Delinquency scale to be related to the disciplinary incidents criterion, yet the empirically-based approach did not support its inclusion in that screen. These results suggest that some of the findings may be due to spurious or suppression effects rather than a reflection of actual explanatory relationships between the predictor scales and the criterion. For this reason, combined theoretical-empirically derived screens were developed as a comparison.

To derive the combined theoretical-empirical screens, the definitions of the TAPAS scales were reviewed and linked to individual criteria. These theoretical relationships were also supported by the bivariate correlations reported in Table 5.8. A hierarchical regression with forward selection was then used to determine if any of the holdover scales could significantly increment the prediction of each of the five targeted criteria beyond the scales identified based on their hypothesized relations to the criterion. The results of these incremental validity analyses are presented in Table 5.9. Results suggested that two TAPAS scales, Intellectual Efficiency and Order, could be added to the disciplinary incidents screen. The sign of the bivariate correlations was used to determine how a scale was weighted in each screen. Table 5.7 summarizes the composition of the combined theoretical-empirical screens.

Table 5.7. TAPAS Scales Included in Empirically and Theoretically Derived "Best Bet" Composites

TAPAS Scale	Theoretically-Derived Composite					Empirically-Derived Composite				
	JKT	Effort	APFT	Discipline	Attrition	JKT	Effort	APFT	Discipline	Attrition
Achievement	+	+	+		+	+	+	+		
Curiosity	+									
Non-Delinquency				+				–		
Dominance		–								–
Even-Temper				+	+	+			+	+
Attention-Seeking				–		+	–	–	–	
Intellectual Efficiency	+	–		–			–		–	
Order				+				+	+	
Physical Conditioning			+		+	–	+	+		+
Tolerance										
Cooperation/Trust										
Optimism	+	+			+					+

Note. Values in cells indicate whether scale was included in the Empirical or Theoretical "Best Bet" composite. A plus (+) sign indicates the TAPAS scale was positively related to the criterion (e.g., higher JKT score, lower Attrition) and had a positive sign in computation of the composite score. A minus (–) sign indicates the TAPAS scale was negatively related to the criterion (e.g., lower JKT score, higher Attrition) and had a negative sign in the computation of the composite score. JKT = Job Knowledge Test, APFT = Army Physical Fitness Test score

Table 5.8. Bivariate Correlations between Individual TAPAS Scales Key Criteria

	Criterion Measure				
TAPAS Scale	JKT	Effort	APFT	Disc	Attrition
Achievement	.08	.08	**.12**	.01	–.04
Curiosity	**.11**	–.02	.03	–.07	–.04
Non-Delinquency	.05	.03	–.07	**–.09**	.01
Dominance	.05	–.06	–.03	.06	.03
Even-Temper	**.14**	.07	–.05	**–.12**	**–.10**
Attention-Seeking	.02	–.08	–.03	**.15**	.04
Intellectual Efficiency	**.14**	–.02	.00	.02	–.03
Order	.00	–.01	.08	**–.12**	-.01
Physical Conditioning	–.04	.07	**.24**	.00	**–.10**
Tolerance	–.01	–.03	–.04	–.04	-.02
Cooperation/Trust	–.03	.04	–.06	–.04	–.02
Optimism	**.15**	.03	–.02	.03	**–.10**

Note. n = 505-1,696. Statistically significant correlations are bolded, $p < .05$ (two-tailed). JKT = Job Knowledge Test, Effort = AW PRS measure of overall Effort, APFT = Army Physical Fitness Test, Disc = Disciplinary incidents during IET, Attrit = 6-month attrition.

Table 5.10 summarizes the intercorrelations among the empirically-based and the combined theoretical-empirical derived TAPAS screens. There are two findings of note. First, the intercorrelations among the screens within the same approach were generally low, with most being below the .30s. This finding supports the creation and use of multiple performance screens, each geared towards a specific criterion. The second finding of note is that the correlations between screens for corresponding criteria across the two different approaches ranged from .28 (the JKT screen) to .93 (the disciplinary incidents screen), indicating the two sets of screens were not redundant with each other.

Table 5.9. Incremental Validity Estimates of Combined Theoretical-Empirical "Best Bet" TAPAS Scales over the AFQT for Predicting Selected Criteria

Criterion	*n*	Step 1 AFQT Only	Step 2 Selected TAPAS	Step 3 Remaining TAPAS	Δ*R* Over AFQT	Δ*R* Over Selected TAPAS
Job-Specific Knowledge	504	**.476**	.487	N/A	**.011**	N/A
AW Effort	538	**.184**	**.236**	N/A	**.052**	N/A
Army Physical Fitness Test (APFT)	521	.035	**.258**	N/A	**.222**	N/A
Disciplinary Incidents	522	.036	**.253**	N/A	**.218**	N/A
6-Month Attrition	1,689	.049	**.230**	N/A	**.182**	N/A

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. *Step 1 AFQT Only* = Correlation between the AFQT and the criterion. *Step 2 Selected TAPAS* = Multiple correlation (*R*) between the AFQT + Selected TAPAS scales with the criterion. *Step 3 Remaining TAPAS* = *R* between AFQT + Selected TAPAS scales + scales included with the forward selection method. Δ*R Over AFQT* = Increment in *R* over the AFQT from adding the selected TAPAS scales to the regression model. Δ*R Over Selected TAPAS* = Increment in *R* over AFQT and selected TAPAS scales from adding additional scales with forward selection. The *R*'s for Disciplinary Incidents and 6-Month Attrition were estimated using Nagelkerke's method in logistic regression. Estimates in bold are statistically significant, $p < .05$ (two-tailed).

Evaluation of TAPAS-Based Performance Screens

Tables 5.6 and 5.9 also provide the incremental validity results for the combined theoretical-empirical and purely empirical results. The results of this analysis show a clear preference for the empirically-derived composite screens. The incremental validity for the empirically-derived screens ranged from ΔR = .02 to .25 (average ΔR = .15) and from ΔR = .01 to .22 (average ΔR = .12) for the combined theoretical-empirical screens. However, since the empirical TAPAS screens were derived from regression results, they were more susceptible to spurious covariation than the theoretical composites. In other words, the empirical composites were more likely to appear stronger in the regression results because they were derived from regression-based analyses. When conceptualized in this way, the question then becomes whether any explanatory power would be lost by using a combined theoretically-empirically derived screen instead of an empirically-derived one.

For this reason, split-group analyses, which more closely approximate how the TAPAS screens would be used operationally, were conducted to evaluate the two types of composites. The results of these analyses are reported in Tables 5.11 and 5.12. The key comparisons of interest in both tables are (a) the difference between the criterion scores for the top third of AFQT Category IIIB scores and the Category IIIA scores and (b) the difference between the top third of Category IIIB scores and the middle and bottom third of the Category IIIB scores. The first comparison is reflective of an approach that could be used to "screen in" Tier 1 Category IIIB accessions during a difficult recruiting period. The second comparison reflects an approach to "screen out" less motivated applicants during an affluent recruiting period. The optimal results have positive *d*-coefficients for continuous criteria and Relative Risk ratios below 1.0 for dichotomous criteria when comparing the top third Category IIIB score and the Category IIIA, scores. The top third Category IIIB scores should also be higher than the middle and bottom third scores. Overall, results support the use of the TAPAS in new Soldier selection. Category IIIB Soldiers that have the highest TAPAS-derived composite scores (i.e., top third) also tended to have higher APFT scores, higher or comparable AW PRS Effort scores, lower rates of attrition, and fewer disciplinary incidents than Category IIIA Soldiers. These results are most striking for the two dichotomous criteria, where Category IIIA Soldiers were 40% to 60% more likely to have an incident than "screened in" Category IIIB Soldiers. The lone exception to this pattern was for the MOS-specific JKT. Mean scores for Category IIIB Soldiers selected in with the JKT TAPAS composite tended to perform worse on the JKT than Soldiers in Category IIIA. However, Soldiers with the top third of scores for the JKT outperformed other Category IIIB Soldiers, suggesting the screen was operating as expected for this group as well.

In comparing the combined theoretical-empirical and empirically-derived screens on the split-group analyses, results were fairly comparable for most cases. However, the results tended to favor the theoretically-derived composites for the two dichotomous criteria. In other words, the likelihood of having an incident (attriting or discipline) was lower for Soldiers "selected in" with the theoretical screen than with the empirical screen. For two of the three continuous criteria, the results were almost exactly the same for the two methods. The mean scores for Soldiers "selected in" using either the theoretical or empirical APFT and Effort composites were roughly equal. Finally, results for the JKT composites tended to favor the empirical approach over the theoretical. Mean JKT scores for Soldiers selected in with the empirical composite were

higher than for Soldiers selected in with the theoretical composite. Overall, these results suggest that for four of the five criteria, little is lost by using the combined theoretical-empirical composite screens over the empirical-only screens.

Table 5.10. Means, Standard Deviations, and Intercorrelations of the Empirically-Derived and Combination TAPAS Composites

Composite	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Combination – JKT	-0.19	1.68									
2. Combination – PRS Effort	0.42	1.12	**.08**								
3. Combination – APFT	0.29	1.03	**.41**	**.32**							
4. Combination – Disciplinary Incidents	-0.09	1.73	.01	**.30**	.03						
5. Combination – 6-Month Attrition	-0.27	1.53	**.61**	**.43**	**.72**	**.27**					
6. Empirical – JKT	-0.59	1.28	**.28**	**.13**	**-.07**	-.01	**.27**				
7. Empirical – PRS Effort	0.60	1.37	**.07**	**.53**	**.70**	**.51**	**.47**	**-.36**			
8. Empirical – APFT	0.26	1.50	**.35**	**.25**	**.73**	**.26**	**.53**	**-.27**	**.72**		
9. Empirical – Disciplinary Incidents	-0.21	1.40	**-.04**	**.34**	.03	**.93**	**.30**	.00	**.53**	**.41**	
10. Empirical – 6-Month Attrition	-0.29	1.42	**.32**	**.49**	**.40**	**.28**	**.81**	**.10**	**.30**	**.31**	**.35**

Note. n = 3,381. Statistically significant correlations are bolded, $p < .05$ (two-tailed). Combination = TAPAS composites computed using a combination of theoretical rationale and empirical data. Empirical = TAPAS composites computed using best subsets analysis. JKT = Job Knowledge Test, PRS Effort = AW PRS measure of overall Effort, APFT = Army Physical Fitness Test, Disciplinary Incidents = Disciplinary incidents during IET (Yes/No). Dashed boxes indicate intercorrelations among the TAPAS composites derived using the same method (e.g., empirical).

Table 5.11. Split Group Analysis Using Unit-Weighted "Best Bet" TAPAS Composites for Predicting Continuous Criteria

Predictor	AFQT Category												
	I-IIA		IIIA		Top 1/3rd IIIB		IIIB-IIIA	Middle 1/3 IIIB		IIIB-IIIA	Bottom 1/3 IIIB		IIIB-IIIA
	M	*SD*	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*d*	*M*	*SD*	*d*
Theoretically-Derived Composites													
Job Knowledge	0.58	0.84	-0.12	0.88	-0.29	0.98	-0.19	-0.30	0.97	-0.20	-0.55	0.95	-0.48
PRS Effort	3.69	0.63	3.50	0.71	3.58	0.78	0.12	3.44	0.58	-0.08	3.34	0.76	-0.22
APFT	242.41	31.85	242.87	29.40	248.78	30.14	0.20	236.68	35.65	-0.21	235.06	27.09	-0.27
Empirically-Derived Composites													
Job Knowledge	0.58	0.84	-0.12	0.88	-0.19	0.98	-0.07	-0.33	0.95	-0.23	-0.69	0.92	-0.64
PRS Effort	3.69	0.63	3.50	0.71	3.47	0.74	-0.04	3.52	0.72	0.03	3.38	0.75	-0.17
APFT	242.41	31.85	242.87	29.40	247.12	32.92	0.14	238.97	32.19	-0.13	236.13	29.41	-0.23

Note. M and *SD* reflect the scores for the dependent variable of interest. Cohen's $d = [(M_{IIIA} - M_{IIIB}) / SD_{IIIA}]$. Category IIIB Soldiers were split into thirds based on a unit-weighted composite of the "best bet" predictor composites for the dependent variable of interest. PRS Effort = AW PRS measure of overall Effort, APFT = Army Physical Fitness Test.

Table 5.12. Split Group Analysis Using Unit-Weighted "Best Bet" TAPAS Composites for Predicting Dichotomous Criteria

Predictor	AFQT Category												
	I-IIA		IIIA		Top 1/3rd IIIB		IIIB-IIIA	Middle 1/3 IIIB		IIIB-IIIA	Bottom 1/3 IIIB		IIIB-IIIA
	$\%_{INCDNT}$	*n*	$\%_{INCDNT}$	*N*	$\%_{INCDNT}$	*n*	*RR*	$\%_{INCDNT}$	*n*	*RR*	$\%_{INCDNT}$	*n*	*RR*
Theoretically-Derived Composites													
Disciplinary Incident	24.8	165	29.9	154	18.6	59	0.62	25.8	62	0.86	37.7	69	1.26
6-Month Attrition	9.2	672	12.5	375	8.3	180	0.66	8.4	178	0.67	19.8	182	1.58
Empirically-Derived Composites													
Disciplinary Incident	24.0	371	32.1	315	18.9	53	0.59	25.4	71	0.79	37.9	66	1.18
6-Month Attrition	9.2	672	12.5	375	8.9	180	0.71	10.1	178	0.81	17.6	182	1.40

Note. $\%_{INCDNT}$ = % of Soldiers in group, out of total *N*, that had an incident (either attriting within 6 months or having at least one disciplinary incident before the end of training). RR = Relative Risk Ratio (*p*[IIIB subgroup attriting or having a disciplinary incident]/*p*[IIIA attriting or having a disciplinary incident]). Category IIIB Soldiers were split into thirds based on a unit-weighted composite of the two types of "best bet" predictors.

Development and Evaluation of a Performance Screen for IOT&E

Approach

The previous sections covered the selection of the experimental predictor measure(s), of which the TAPAS emerged as the temperament measure having the highest potential for enhancing new Soldier selection. In this section, we summarize the development and evaluation of the performance screen to be used in the IOT&E. The goal in these analyses was to select a single, performance screen that could be used to identify Tier 1, AFQT Category IIIB applicants likely to perform similarly to Category IIIA Soldiers.

The intent of this screen was to provide a basis for increasing the number of Tier 1 Category IIIB accessions during a difficult recruiting period. However, due to sudden changes in the recruiting market, driven by significant and unexpected increases in unemployment, the screen developed herein was not implemented in the IOT&E (see Chapter 6 for further information on the screen that is being used instead). During the IOT&E the Army will continue to evaluate the potential of this screen as a market expansion tool for possible future use as market conditions or priorities change.

Developing a Performance Screen for IOT&E

Decades of research stemming from the Army's Project A (Campbell & Knapp, 2001) reinforces the value in treating job performance as a multidimensional construct. At the broadest level, Project A distinguished between two major components of performance; "can-do" and "will-do." The can-do component typically consists of performance dimensions that reflect technical competence. Conversely, the will-do component consists of performance dimensions such as Effort and Leadership and Maintaining Personal Discipline.

Two considerations guided the development and evaluation of the performance screen for the IOT&E: (a) to select TAPAS scales that would significantly enhance the prediction of can-do performance dimensions over cognitive ability, and (b) to select TAPAS scales that could also identify Category IIIB applicants that are highly motivated and likely to excel at critical will-do performance dimensions. Research suggests that matching narrowly defined performance criteria with predictors results in greater predictive validity (Tett, Jackson, & Rothstein, 1991). Further, technical or can-do performance tends to be more strongly predicted by cognitive ability and less by temperament and other noncognitive characteristics, whereas the latter characteristics are generally considered to be strong predictors of will-do performance (Campbell & Knapp, 2001; Motowidlo & Van Scotter, 1994; Van Scotter & Motowidlo, 1996).[8] For these reasons, the development of potential performance screens started with dividing the criterion space into can-do and will-do performance dimensions to identify the criteria to be used in their development.

The identification and selection of these criteria involved the following steps. First, the performance criteria available at the time of analyses were categorized by military personnel researchers into can-do or will-do performance dimensions. Next, the researchers chose specific, targeted criteria for best representing the can-do and will-do dimensions. Three criteria were

[8] An exception to this is Conscientiousness (see Hurtz & Donovan, 2000, for a review), which has also been found to be a strong predictor of can-do dimensions of performance.

selected for representing the can-do domain (i.e., MOS-specific JKT, average AIT exam grade, and graduation from AIT/OSUT) and seven criteria were chosen to measure the will-do domain (i.e., APFT score, disciplinary incidents, self-rated Adjustment to Army Life, PRS scales of Effort, Discipline, and Peer Support, and 6-month attrition).

To create the two noncognitive-focused composites (can-do and will-do) for use in the performance screen, we evaluated the TAPAS scales using a combination of theoretical insight and empirical evidence that included an examination of (a) their correlations with targeted criteria, (b) the strength of these relationships, and (c) their overall pattern of correlations across the different criteria (e.g., is a negative correlation justified by theory and previous research?). Regression analyses were also run for each criterion, controlling for AFQT scores. Ultimately, however, the zero-order correlations and theoretical rationale were used to choose the TAPAS scales with the greatest potential for predicting can-do and will-do performance. The decision to rely on rational considerations and zero-order correlations was made based on a concern with over-reliance on the research sample, small sample sizes, and the diversity of criteria and resulting outcomes. Such an approach is also consistent with the results from the preceding section. Those results demonstrated that there would be no measurable loss in predictive validity from using a screen based on a combined theoretical-empirical approach than one constructed using a purely empirically-based approach.

Once the TAPAS scales were selected they were combined into two unit-weighted composites focused on can-do and will-do performance, respectively. The resulting composites were correlated with the targeted criteria to examine their predictive efficacy. Next, several different ways of combining these two composites were considered to best identify the Category IIIB applicants likely to perform more like Category IIIA Soldiers (e.g., averaging scores on scales constituting each composite, combining average scores on the two composites, multiple hurdle). Ultimately, due to greater predictive strength, the decision was made to calculate the two composites separately and use a multiple hurdle approach where individuals have to pass both the can-do and will-do composites to obtain a passing score on the performance screen. The minimum passing score was set at the 50th percentile on both TAPAS composites. In addition, we restricted the Category IIIB applicants who could pass the screen to those with AFQT scores between 40 and 49. This range of AFQT scores was selected because they represent the top 50% of Category IIIB applicants, and are therefore more likely to perform similarly to Category IIIA Soldiers.

In summary, a final performance screen for the IOT&E, called the Tier One Performance Screen [TOPS], was developed for identifying Educational Tier 1, Category IIIB Soldiers most likely to perform similarly to Category IIIA Soldiers. It was determined that in order to be identified as a "high potential " Category IIIB applicant, the individual must meet the following criteria: (a) AFQT score between 40-49, (b) score in top 50th percentile of TAPAS can-do composite, and (c) score in top 50th percentile of noncognitive will-do composite. Soldiers meeting these criteria were classified as "passing TOPS," while those not meeting these criteria were classified as "failing TOPS."

Evaluating the Performance Screen for IOT&E

The resulting TOPS screen was then evaluated for use as an operational, applicant screen using two metrics. The first was based on the "split-group" technique described earlier in this

chapter. Specifically, we compared the performance of Tier 1 Category IIIB Soldiers, passing the TOPS, as defined above, with the performance of (a) Category I-IIIA Soldiers, and (b) Category IIIB Soldiers failing TOPS. Same as with the split group analyses presented earlier, the groups were compared using Cohen's *d* for continuous criteria, and a Relative Risk (RR) ratio for the dichotomous criteria.

The second metric by which the final TOPS screen was evaluated was its adverse impact. Adverse impact ratios were calculated for gender and race. Calculating the adverse impact ratio consisted of two steps. First, the selection rates were calculated for each gender x race grouping separately (e.g., White females, White males, Black females, Black males, Hispanic females, and Hispanic males). Second, the rates were then weighted by the percentage of (a) males or females for the ratios evaluating racial adverse impact and (b) White, Blacks, and Hispanics for the ratios evaluating gender adverse impact. These steps were taken to ensure that adverse impact ratios for gender were not contaminated by race and vice versa.

Findings

Developing the Performance Screen for IOT&E

The zero-order correlations between all TAPAS scales and targeted criteria are presented in Tables 5.13 and 5.14. Table 5.13 provides the correlations between the TAPAS scales and the "can-do" criteria, while Table 5.14 shows the correlations with the targeted "will-do" criteria.

Based on examination of correlations of individual TAPAS scales with the targeted criteria, five scales were chosen as good indicators of "can-do" performance potential (i.e. Achievement, Non-Delinquency, Even-Temper, Intellectual Efficiency, and Optimism) and five scales were chosen as good indicators of "will-do" performance potential (i.e. Achievement, Non-Delinquency, Even-Temper, Attention-Seeking, and Physical Conditioning). These scales were combined into two separate, unit-weighted composites—one for prediction of can-do performance criteria and one for prediction of will-do performance criteria. The correlation between these two composites was .73 ($p < .05$).

The correlations of the can-do and will-do composites with the targeted performance measures are presented at the bottom of Tables 5.13 and 5.14. The can-do composite is a better predictor of the can-do criteria, while the will-do composite is a significant predictor across all of the will-do criteria.

Evaluating the Performance Screen for IOT&E

Table 5.15 presents the results of the split-group analyses for evaluating the TOPS performance screen. Overall, the results show that Tier 1, Category IIIB Soldiers who pass TOPS perform similarly to Category IIIA Soldiers. The differences in mean criterion scores between Category IIIB Soldiers passing TOPS and Category IIIA Soldiers are either in the positive direction (i.e., Category IIIB Soldiers passing TOPS perform better than Category IIIA Soldiers) or very minimally in the negative direction. Additionally, Category IIIB Soldiers passing TOPS perform better across criteria than those that fail TOPS. This provides evidence that TOPS may be a useful screening tool for Category IIIB applicants.

As shown in Table 5.16, no evidence of adverse impact was found. In fact, female and Hispanic Soldiers had higher pass rates than males and White Soldiers, respectively. Black Soldiers had slightly lower pass rates on TOPS than White Soldiers, but not enough to indicate significant adverse impact. In comparison to the current composition of the Army, implementing the TOPS as an applicant screening measure would be expected to increase the number of Black, Hispanic, and female Soldiers entering the Army.

Table 5.13. Correlations between Predictor Measures (AFQT and TAPAS-95s Scales) and Targeted "Can-Do" Criteria

		Targeted Can-Do Criteria	
Predictor Measure/Scale	Job Knowledge Test	Average AIT Exam Grade	Graduation from AIT/OSUT
AFQT	**.48**	**.34**	**.04**
TAPAS-95s Scales			
Achievement	.08	**.18**	.02
Curiosity	**.11**	**.09**	.02
Non-Delinquency	.06	**.13**	.01
Dominance	.05	.05	-.03
Even-Temper	**.14**	**.12**	**.09**
Attention-Seeking	.02	-.05	-.03
Intellectual Efficiency	**.14**	**.15**	.00
Order	.01	.05	-.02
Physical Conditioning	-.04	-.03	**.08**
Tolerance	-.01	.04	.00
Cooperation/Trust	-.03	.01	.02
Optimism	**.15**	.08	**.08**
TAPAS-95s Composites			
Can-Do	**.22**	**.25**	**.07**
Will-Do	.08	**.17**	**.09**

Note. n = 505 – 2,535. Statistically significant correlations are bolded, $p < .05$ (two-tailed). Correlations are uncorrected for statistical artifacts (e.g., range restriction).

Table 5.14. Correlations between Predictor Measures (AFQT and TAPAS-95s Scales) and Targeted "Will-do" Criteria

	Targeted Will-Do Criteria						
Predictor Measure/Scale	Army Physical Fitness Test	Disciplinary Incident (Y/N)	Adjustment to Army Life	Ratings of Effort	Ratings of Discipline	Ratings of Peer Support	6-Month Attrition
AFQT	.04	**-.09**	**.13**	**.18**	**.17**	**.15**	-.02
TAPAS-95s Scales							
Achievement	**.12**	.00	**.24**	.08	.05	-.02	-.04
Curiosity	.03	-.04	.07	-.02	-.01	-.02	-.04
Non-Delinquency	-.07	**-.11**	.05	.03	**.16**	.08	.01
Dominance	-.03	.03	.07	-.06	**-.12**	.07	.03
Even-Temper	-.05	**-.09**	.08	.07	**.15**	.07	**-.10**
Attention-Seeking	-.03	**.14**	.01	-.08	**-.17**	**-.09**	.04
Intellectual Efficiency	.00	.07	.04	-.03	-.05	-.05	-.03
Order	.08	**-.10**	.04	-.01	.03	-.04	-.01
Physical Conditioning	**.25**	-.02	**.19**	.07	.00	.02	**-.10**
Tolerance	-.04	.01	**.11**	-.03	.00	.04	-.02
Cooperation/Trust	-.06	-.07	-.06	.04	**.11**	.04	-.02
Optimism	-.03	.02	**.14**	.03	.02	.02	**-.10**
TAPAS-95s Composites							
Can-Do	-.02	-.05	**.22**	.07	**.14**	.04	**-.10**
Will-Do	**.10**	**-.14**	**.20**	**.13**	**.21**	**.10**	**-.11**

Note. n = 505 – 1,696. Statistically significant correlations are bolded, $p < .05$ (two-tailed). Correlations are uncorrected for statistical artifacts (e.g., range restriction).

Table 5.15. Split Group Analysis Comparing Soldiers by AFQT Category on Targeted Continuous and Dichotomous Criteria

	AFQT Category – Continuous Criteria									
	I-II		IIIA		IIIB Pass TOPS		IIIB-IIIA	IIIB Fail TOPS		IIIB-IIIA
Criterion	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*d*
Job Knowledge Test	94.35	22.81	74.07	26.33	71.61	20.93	-.09	64.99	28.20	-.34
Course Average	96.13	6.80	93.64	6.76	93.33	4.19	-.05	90.17	6.98	-.51
Army Physical Fitness Test	246.01	32.10	243.31	30.04	244.54	41.95	.04	239.44	29.66	-.13
Adjustment to Army Life	3.83	.63	3.64	.68	3.82	.59	.26	3.62	.72	-.03
Ratings of Effort	3.72	.65	3.57	.66	3.63	.72	.09	3.44	.73	-.20
Ratings of Discipline	3.97	.65	3.86	.64	3.80	.72	-.09	3.64	.69	-.34
Ratings of Peer Support	3.89	.59	3.82	.59	3.77	.60	-.08	3.65	.63	-.34

	AFQT Category – Dichotomous Criteria									
	I-II		IIIA		IIIB Pass TOPS		IIIB-IIIA	IIIB Fail TOPS		IIIB-IIIA
	$\%_{INCDNT}$	*N*	$\%_{INCDNT}$	*N*	$\%_{INCDNT}$	*N*	*RR*	$\%_{INCDNT}$	*N*	*RR*
Graduation from Training	91.4	1,653	89.3	1,273	95.0	141	1.06	88.5	794	.99
Disciplinary Incident	24.0	371	32.1	315	23.1	26	.72	28.7	164	.89
6-Month Attrition	9.3	1,127	10.6	727	7.6	79	.72	13.0	461	1.23

Note. IIIB-IIIA d = standardized mean difference in criterion scores (or Cohen's *d*) between the selected IIIB and IIIA Soldiers [(M_{IIIB} - M_{IIIA}) / SD_{IIIA}]. $\%_{INCDNT}$ = % of Soldiers, out of the total number (*N*), in selected AFQT Category that exhibited an incident on the criterion measure (e.g., attriting within 6 months). *IIIB-IIIA RR* = relative risk ratio of selected IIIB Soldiers having an incident relative to that of IIIA Soldiers (*p*[IIIB having an incident]/*p*[IIIA having an incident]).

Table 5.16. Adverse Impact Ratios for TOPS

	Adverse Impact Ratios for Race					
	Female Pass Rates		Male Pass Rates			
	Actual	Weighted	Actual	Weighted	Overall Pass Rate	Adverse Impact Ratio
Black	.177	4.66	.121	8.92	13.57	.93
Hispanics	.333	8.76	.136	10.02	18.78	1.29
Whites	.184	4.84	.132	9.73	14.57	

	Adverse Impact Ratio for Gender							
	Black Pass Rates		Hispanic Pass Rate		White Pass Rates			
	Actual	Weighted	Actual	Weighted	Actual	Weighted	Overall Pass Rate	Adverse Impact Ratio
Females	.177	3.61	.333	6.06	.184	11.30	20.97	1.61
Males	.121	2.47	.136	2.48	.132	8.10	13.05	

Note. Adverse impact ratios above 1 indicate that the minority group passes TOPS at a greater rate than the majority group. Adverse impact ratios under .80 represent adverse impact. The composition of gender and race used to compute the weighted pass rates are as follows: Female – 26.3%, Male – 73.7%, Black – 20.4%, Hispanic – 18.2%, White – 61.4%.

CHAPTER 6: CONCLUSIONS AND NEXT STEPS

Tonia S. Heffner and Leonard White (ARI)

The results from the *Expanded Enlistment Eligibility Metrics (EEEM)* research and analyses demonstrate that non-cognitive measures can assess a Soldier's potential and, in combination with AFQT, predict training performance and attrition. In particular, non-cognitive measures incrementally predict "will-do" criteria such as attrition, physical fitness, and disciplinary incidents as well as combine with AFQT to improve the prediction of "can-do" criteria such as academic performance. Further, non-cognitive measures have the potential to increase diversity within the Army.

To identify particular non-cognitive measure(s) for use in an enlistment screen, we evaluated estimated incremental validities over AFQT, subgroup differences, potential for faking, and overall administration requirements. All of the temperament measures (TAPAS, RBI, AIM) demonstrated strong incremental validity over AFQT. The TAPAS and AIM had quite low subgroup differences and all three measures had scales that favored minority groups over the majority. Additionally, all three measures have equivalent administration times since the number of scales can be modified to meet the operational constraints. The TAPAS was selected for use in the selection screen because the research suggests it has the lowest potential for faking in an operational environment.

At the outset of the EEEM research effort, the emphasis was on a supplemental assessment that would permit the Army to "screen in" high potential applicants. Specifically, the goal was to identify applicants whose scores on the screening measures suggest that they have the motivation and potential to perform at a higher level than their AFQT scores would predict. These results clearly demonstrate that applicants categorized as IIIB, but scoring high on the TAPAS, would perform like IIIA Soldiers (see Tables 5.11, 5.12, 5.15). This is significant because Category I-IIIA applicants are eligible for a more diverse set of jobs and enlistment incentives; thus TAPAS could function as a market expander for the Army. During the short course of this research, however, the recruiting market changed dramatically so that the number of applicants exceeded the number of Soldiers needed. Instead of "screening in" high potential candidates, the emphasis became "screening out" applicants with the lowest motivation and potential.

The inherent flexibility in the TAPAS measure and the thoroughness of the data analysis allowed ARI to quickly adapt the TOPS program to the new mission requirements. Specifically, the TAPAS composite scores were compared for applicants in the lowest acceptable cognitive ability category, Category IV, as determined by their AFQT scores. These results demonstrated that the TAPAS composites were equally appropriate for this AFQT category as for the other categories. The TAPAS composites were demonstrated to discriminate between high from low performing Soldiers (see Table 6.1). As can be seen in Table 6.1, AFQT Category IV Soldiers passing TAPAS, as compared with those who do not pass, had higher training graduation rates, higher AIT exam grades, and a 50% lower attrition rates at 6 months of Service.

Table 6.1. Effects of TAPAS Screening for Category IV Soldiers

	Percent Passing TAPAS		Percent Failing TAPAS	
Criteria	Mean	*SD*	Mean	*SD*
AIT Exam Grades	88.71	7.62	83.11	26.18
6-month Attrition	4.00		12.00	
Training Graduation	94.00		92.00	
Training Recycles	15.00		13.00	

Note. n = 12 - 108. Sample includes only Educational Tier 1 non-prior service Soldiers. No standard deviation provided for the three dichotomous criteria.

The results also support continued research on the WPA as a supplement to the ASVAB as a classification tool. The WPA had significant incremental validities for the MOS-specific performance-related and retention-related criteria, suggesting that MOS fit may play a role in Soldiers' overall success in the Army.

U.S. Army Implementation

The EEEM results were sufficiently encouraging to garner the support of Army leadership to move forward with a 3-year initial operational test and evaluation (IOT&E) of the computer adaptive version of the Tailored Adaptive Personality Assessment System (TAPAS) to supplement the AFQT as a selection tool. The IOT&E, officially called the *Tier One Performance Screen (TOPS)*, allows for a longitudinal validation of the non-cognitive measures in a high-stakes testing environment.

The Army transitioned the TAPAS into applicant testing locations between May and August 2009. It now is administered to all Army applicants who take the computer adaptive ASVAB in the Military Entrance Processing Stations (MEPS). Approximately 60% of Army applicants, or 180,000 per year, take the computer adaptive ASVAB whereas the remainder take a paper and pencil version at another location before going to the MEPS. For Category IV applicants, their TAPAS scores determine if they are eligible to enter the Army. Those applicants who score below the 10th percentile on either the "can-do" or "will-do" composites are not enlistment eligible. By design, this screen only impacts a small portion of applicants, but as ASVAB scores are unknown to the applicant when taking the TAPAS, the applicant does not know the impact on his or her eligibility until both tests are completed. Depending on the Army's goals and the applicants' overall educational, medical, moral, educational, cognitive aptitude, as well as the TAPAS assessments, approximately one-third to one-half of the Category IV applicants will be accepted into the Army.

High Stakes Assessment of the Work Preferences Assessment (WPA)

The WPA was selected for additional research and development because of the strength of the research findings and the possible unique information for MOS classification that it can contribute to the TOPS evaluation. Although it will not be used as an operational test at this time, a slightly revised version will be administered for research purposes in the MEPS. Because of the

intense time constraints in the MEPS, the original version of the WPA was evaluated to be too lengthy for administration in that context. Analyses were conducted to reduce the number of items without sacrificing the psychometric properties. Results of these analyses are reported in Appendix E. The high stakes testing conducted in the MEPS should provide a quasi-operational environment in which to better evaluate the WPA as a classification tool.

Tier One Performance Screen Validation

The key feature of the TOPS program is the 3-year longitudinal validation. According to Army directive, the TAPAS will be administered to the majority of Educational Tier 1 applicants who are accessed into the Army. To be an operational eligibility enlistment screen, a measure must screen out some number of applicants. By design, this is limited to a very small number of applicants. This procedure will allow us to track Soldiers to the completion of training and evaluate the TAPAS composite prediction of Army outcomes in a context which limited range restriction in the TAPAS predictor scales and composite measures can be expected. The TAPAS data from the IOT&E will be used to refine the can-do and will-do composites, and validate a variety of approaches for using TAPAS to help the Army to meet its annual accession mission requirements and recruit a high potential force. It will also allow us to evaluate the WPA as a prospective measure to supplement the ASVAB/TAPAS combination as an MOS classification determinant.

The operational validation will mirror the research validation (Knapp & Heffner, 2009). All Soldiers, regardless of whether they took the TAPAS, will be tracked through training and into their first unit of assignment. The TOPS IOT&E training criterion measures consist of performance, retention, and attitudinal variables. For all Soldiers, data will be collected from Army databases as available. This data will include attrition and academic performance. As in the Army Class/EEEM research, Soldiers in the jobs of Infantryman, Armor Crewman, Military Police, Wheeled-Vehicle Mechanic, Medic, and Truck Driver, will be administered an expanded set of criterion measures. Soldiers in the jobs of Signal Support Specialist and Human Resources Specialist also will be added to this list to maximize evaluation of the WPA for classification. The targeted Soldiers will complete Army-wide and MOS-specific job knowledge tests and the ALQ. Their Drill Sergeants, Cadre, or Instructors will provide Army-wide and MOS-specific performance ratings. The number of Soldiers in the 3-year IOT&E could exceed 100,000 although for most Soldiers this data will be limited to archival records. For the targeted Soldiers, the sample could reach 30,000. Because of the enormity of this data collection, the Army has directed the Drill Sergeants/Cadre/Instructors to proctor the data collection as part of the training completion activities with the continued training and support of ARI.

Potential Uses for Non-Cognitive Assessment

A notable advantage of the TOPS is its inherent flexibility. Depending on mission requirements, it can be used to screen out applicants who are unlikely to perform to standard or it can be used to screen in applicants who are likely to perform better than their AFQT scores would predict. The TOPS measures are under consideration for other Army applications – both initial selection and in-service selection. For initial selection, the data collection in the TOPS IOT&E has been expanded to include non-high school diploma, or Educational Tier 2, Soldiers.

Currently, Tier 2 Soldiers are screened by the Tier Two Attrition Screen (TTAS); which has the Assessment of Individual Motivation (AIM) as one of the components. The TTAS is used to evaluate attrition risk in Tier 2 applicants, identifying those in this group who are likely to have attrition rates closer to Tier 1 applicants. In the course of this IOT&E, we are planning to examine how TAPAS can be used in combination with AIM/TTAS to identify applicants with a lower attrition risk. One possible outcome of this research is that TAPAS can replace AIM in the TTAS screen without sacrificing, and possibly improving, attrition prediction. The TAPAS also may be able to better screen applicants for MOS that are hard-to-fill or those that have higher attrition rates.

REFERENCES

Allen, M.T., Cheng, Y.A., Ingerick, M.J., & Caramagno, J.P. (2009). In D.J. Knapp & T.S. Heffner (Eds.) *Predicting future force performance (Army Class): End of training longitudinal validation* (pp. 24-30) (Technical Report 1257). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Andrich, D. (1989). A probabilistic IRT model for unfolding preference data. *Applied Psychological Measurement, 13*, 193-216.

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. 3307-333). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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., McHenry, J.J., & Wise, L.L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology*, *43*, 313-333.

Chernyshenko, O.S., Stark, S., Drasgow, F., & Roberts, B.W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*, 88-106.

Drasgow, F. , Stark, S., & Chernyshenko, O.S. (November, 2006). *Toward the Next Generation of Personality Assessment Systems to Support Personnel Selection and Classification Decisions*. Paper presented at the 48th annual conference of the International Military Testing Association, Kingston, Ontario, Canada.

Goldberg, L.R. (1990). An alternative "description of personality": The Big Five factor structure. *Journal of Personality & Social Psychology*, *59*, 1216-1229.

Graham, K.E., McDaniel, M.A., Douglas, E.F., & Snell, A.F. (2002). Biodata validity decay and score inflation with faking: Do item attributes explain variance across items? *Journal of Business and Psychology, 16*, 573-592.

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

Holland, J.L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources, Inc.

Hurtz, G.M., & Donovan, J.J. (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology*, *85*, 869-879.

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.

Kilcullen, R.N., Putka, D.J., McCloy, R.A., & Van Iddekinge, C.H. (2005). Development of the Rational Biodata Inventory. In D.J. Knapp, C.E. Sager, & T.R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (pp. 105-116) (Technical Report 1168). Arlington, 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., Heggestad, E.D., & Young, M.C. (Eds.) (2004). *Understanding and improving the Assessment of Individual Motivation (AIM) in the Army's GED Plus Program* (Study Note 2004-03). Alexandria, 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.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics, 15*(4), 661-675.

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

Moriarty, K.O., Campbell, R.C., Heffner, T.S., & Knapp, D.J. (2009). *Investigations into Army enlisted classification systems (Army Class): Reclassification test and criterion development report* (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.

Press, W.H., Flannery, B.P., Teukolsky, S.A., & Vetterling, W.T. (1990). *Numerical recipes: The art of scientific computing. New* York: Cambridge University Press.

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). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Roberts, J.S., Donoghue, J.R., & Laughlin, J.E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

Russell, T.L., Peterson, N.G., Rosse, R.L., Hatten, J.L.T., McHenry, J. J., & Houston, J.S. (2001). The measurement of cognitive, perceptual and psychomotor abilities. In J.P. Campbell & D.J. Knapp (Eds.) *Exploring the limits in personnel selection and classification* (pp. 71-110). Mahwah, NJ: Lawrence Erlbaum Inc.

Stark, S. (2002). A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment [Doctoral Dissertation]. University of Illinois at Urbana-Champaign.

Stark, S., & Chernyshenko, O.S. (in review). Adaptive testing with the Multi-Unidimensional Pairwise Preference Model. *Applied Psychological Measurement.*

Stark, S., 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., & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement, 26*, 208–227.

Stark, S., Drasgow, F., & Chernyshenko, O.S. (October, 2008). *Update on Tailored Adaptive Personality Assessment System (TAPAS): The Next Generation of Personality Assessment Systems to Support Personnel Selection and Classification Decisions*. Paper presented at the 50th annual conference of the International Military Testing Association. Amsterdam, Netherlands.

Strickland, W.J. (Ed.) (2005). *A longitudinal examination of first term attrition and reenlistment among FY 1999 enlisted accessions* (Technical Report 1172). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Tett, R.P., Jackson, D.N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

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* (pp. 89-104) (Technical Report 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Van Scotter, J. R., & Motowidlo, S. J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology, 81,* 525-531.

Waugh, G.W., & Russell, T.L. (2005). Predictor situational judgment test. In D.J. Knapp & T.R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (pp. 235-154) (Technical Report 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

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.

White, L.A., Young, M.C., Hunter, A.E., & Rumsey, M.G. (2008). Lessons learned in transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*(3), 291-295.

White, L.A., Young, M.C., & Rumsey, M.G. (2001). ABLE implementation issues and related research. In J.P. Campbell & D.J. Knapp's (Eds.), *Exploring the limits of personnel selection and classification* (pp. 525-558). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Zuccaro, C. (1992). Mallow's Cp statistic and model selection in multiple linear regression. *Journal of Market Research Society, 34*, 163-178.

APPENDIX A
DESCRIPTIVE STATISTICS AND SCORE INTERCORRELATIONS FOR SELECTED CRITERION MEASURES

Table A.1. Descriptive Statistics and Reliability Estimates for the Army-Wide (AW) and MOS-Specific Performance Rating Scales (PRS)

	11B Infantryman		19K Armor Crewmen		31B Military Police		63B Light Wheel Vehicle Mechanic		Total EEEM For-Research-Only Criterion Sample				
Composite/Scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*α*	*ICC(A,1)*	*ICC(A,k)*
AW PRS													
Effort Composite	3.60	.74	3.62	.67	3.54	.68	3.57	.73	3.58	.70	.89	.29	.62
Physical Fitness & Bearing Composite	3.95	.71	3.85	.65	3.97	.69	3.86	.65	3.93	.69	.86	.31	.63
Personal Discipline Composite	3.96	.66	3.85	.59	3.79	.70	3.69	.78	3.84	.68	.90	.27	.59
Commitment & Adjustment Composite	3.95	.66	3.75	.69	3.83	.64	3.74	.72	3.83	.67	.84	.24	.55
Support for Peers Composite	3.85	.63	3.79	.57	3.79	.62	3.71	.67	3.80	.62	.85	.16	.43
Peer Leadership Composite	3.44	.80	3.45	.74	3.37	.74	3.57	.73	3.42	.76	.89	.26	.58
Common Warrior Tasks KS Scale	3.92	.71	3.83	.69	4.00	.65	3.87	.63	3.93	.67	n/a	.20	.49
MOS Qualification KS Scale	3.97	.72	3.86	.63	4.02	.59	3.96	.74	3.97	.65	n/a	.17	.45
MOS-Specific PRS Composite [a]	5.09	.86	4.73	.72	5.09	.69	5.23	1.11	5.02	.80	.94	.20	.41

Note. n = 1,158-1,184; 11B Infantryman, n = 290-299. 19K Armor Crewman, n = 254; 31B Military Police, n = 517-532; 63B Light Wheel Vehicle Mechanic, n = 97-99. α = internal consistency reliability estimates (coefficient alpha). *ICC(A,1)* = intraclass correlation coefficient assuming a single rater. *ICC(A,k)* = intraclass correlation coefficient assuming multiple (or *k*) raters. The AW PRS scales range from 1 – 5; the MOS-Specific PRS Composite ranges from 1 – 7.

[a] The reliability estimates for the total sample represent weighted averages across the four MOS; *α* (11B = .96, 19K = .92, 31B = .93, 63B = .96); *ICC(A,1)* (11B = .15, 19K = .19, 31B = .30, 63B = .18); *ICC(A,k)* (11B = .35, 19K = .42, 31B = .49, 63B = .40). Ratings include both peers and supervisors.

Table A.2. Intercorrelations among Army-Wide (AW) and MOS-Specific PRS

	Composite/Scale	1	2	3	4	5	6	7	8
1	AW Effort Composite								
2	AW Physical Fitness & Bearing Composite	.72							
3	AW Personal Discipline Composite	.78	.61						
4	AW Commitment & Adjustment Composite	.76	.74	.78					
5	AW Support for Peers Composite	.72	.60	.78	.75				
6	AW Peer Leadership Composite	.75	.68	.65	.73	.68			
7	AW Common Warrior Tasks KS Scale	.72	.76	.64	.77	.66	.72		
8	MOS Qualification KS Scale	.68	.70	.60	.73	.63	.67	.80	
9	MOS-Specific PRS Composite - Total	.63	.57	.55	.62	.57	.61	.64	.65
9a	MOS-Specific PRS Composite - 11B	.64	.60	.63	.67	.64	.58	.70	.66
9b	MOS-Specific PRS Composite - 19K	.57	.51	.55	.60	.58	.58	.59	.58
9c	MOS-Specific PRS Composite - 31B	.69	.62	.55	.64	.56	.70	.66	.69
9d	MOS-Specific PRS Composite - 63B	.68	.52	.60	.61	.60	.59	.58	.64

Note. n = 1,161-1,187. The correlations between the MOS-specific composite ratings and the AW composites/scales for each MOS are presented in rows 9a through 9d. 11B Infantryman, n = 245; 19K Armor Crewman, n = 254; 31B Military Police, n = 517; and 63B Light Wheel Vehicle Mechanics, n = 97. All correlations are statistically significant, $p < .05$ (two-tailed).

Table A.3. Descriptive Statistics and Reliability Estimates for the Army Life Questionnaire (ALQ) Scales by MOS

	11B Infantryman		19K Armor Crewman		31B Military Police		63B Light Wheel Vehicle Mechanic		Total EEEM For-Research-Only Criterion Sample		
Scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*α*
Commitment and Retention-Related Attitudes											
Attrition Cognitions	4.44	.64	4.39	.61	4.29	.70	4.14	.80	4.34	.68	0.79
Career Intentions	3.32	1.02	3.05	.98	3.00	1.04	3.02	1.06	3.09	1.03	0.94
Army Fit	4.12	.59	4.06	.56	4.02	.60	3.91	.71	4.04	.60	0.81
Affective Commitment	3.99	.65	3.94	.59	3.81	.66	3.70	.71	3.87	.66	0.86
Initial Entry Training (IET) Performance and Adjustment											
Disciplinary Incidents	0.30	.46	0.22	.42	0.33	.47	0.28	.45	0.29	.46	n/a
APFT Score	244.67	31.22	238.39	27.70	246.49	33.20	244.57	32.08	244.05	31.55	n/a

Note. *n* = 1,137-1,149; 11B Infantryman, *n* = 286-289; 19K Armor Crewman, *n* = 247-251; 31B Military Police, *n* = 500-510; 63B Light Wheel Vehicle Mechanic, *n* = 94-95. APFT = Army Physical Fitness Test. α = coefficient alpha. ALQ scale scores range from 1 – 5 except for the following: (a) Disciplinary Incidents (0 = No; 1 = Yes) and APFT Score (free response item, Min = 67, Max = 300).

Table A.4. Intercorrelations among ALQ Scale Scores

Scale		1	2	3	4	5
1	Attrition Cognitions					
2	Career Intentions	**.54**				
3	Army Fit	**.71**	**.54**			
4	Affective Commitment	**.65**	**.56**	**.79**		
5	Disciplinary Incidents	**-.10**	-.05	**-.12**	-.06	
6	APFT Score	**.10**	**.06**	**.15**	.05	**-.10**

Note. *n* = 1,128 -1,141. Statistically significant correlations are bolded, $p < .05$ (two-tailed). APFT = Army Physical Fitness Test.

APPENDIX B
DESCRIPTIVE STATISTICS AND SCORE INTERCORRELATIONS FOR SELECTED PREDICTOR MEASURES

Table B.1. Descriptive Statistics for the Armed Services Vocational Aptitude Battery (ASVAB) Subtests and Armed Forces Qualification Test (AFQT)

Scale	*M*	*SD*
ASVAB Subtests		
General Science (GS)	51.47	7.60
Arithmetic Reasoning (AR)	52.00	6.51
Word Knowledge (WK)	49.79	6.14
Paragraph Comprehension (PC)	51.37	5.19
Math Knowledge (MK)	53.11	6.27
Electronics Information (EI)	51.89	8.04
Auto and Shop Information (AS)	50.24	8.65
Mechanical Comprehension (MC)	53.05	7.78
Assembling Objects (AO)	55.14	7.95
AFQT	57.28	20.15

Note. n = 7,008-8,056. Subtest and composite scores are percentiles.

Table B.2. Intercorrelations among ASVAB Subtest and AFQT Scores

Scale		1	2	3	4	5	6	7	8	9
1	General Science (GS)									
2	Arithmetic Reasoning (AR)	**.41**								
3	Word Knowledge (WK)	**.62**	**.28**							
4	Paragraph Comprehension (PC)	**.44**	**.30**	**.44**						
5	Math Knowledge (MK)	**.31**	**.57**	**.13**	**.18**					
6	Electronics Information (EI)	**.59**	**.39**	**.45**	**.33**	**.20**				
7	Auto and Shop Information (AS)	**.44**	**.27**	**.31**	**.21**	.01	**.59**			
8	Mechanical Comprehension (MC)	**.54**	**.47**	**.38**	**.31**	**.28**	**.60**	**.58**		
9	Assembling Objects (AO)	**.32**	**.40**	**.18**	**.20**	**.33**	**.33**	**.24**	**.50**	
10	AFQT	**.67**	**.77**	**.72**	**.64**	**.67**	**.52**	**.32**	**.54**	**.42**

Note. n = 6,347-7,956. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

Table B.3. Descriptive Statistics and Reliability Estimates for Assessment of Individual Motivation (AIM) Scales

Scale	*M*	*SD*	α
Adjustment	1.25	.29	.74
Agreeableness	1.26	.27	.70
Dependability	1.28	.28	.77
Leadership	1.20	.28	.76
Physical Conditioning	1.20	.34	.78
Work Orientation	1.18	.28	.74
Validity Scale	0.14	.16	n/a

Note. n = 3,286-3,376. α = coefficient alpha. AIM scale scores range from 0 – 2 except for the Validity scale, which ranges from 0 – 1.

Table B.4. Intercorrelations among AIM Scales

Scale		1	2	3	4	5	6
1	Adjustment						
2	Agreeableness	**.63**					
3	Dependability	**.51**	**.51**				
4	Leadership	**.29**	**.16**	**.36**			
5	Physical Conditioning	**.26**	**.26**	**.29**	**.22**		
6	Work Orientation	**.37**	**.29**	**.32**	**.57**	**.52**	
7	Validity Scale	**.13**	**.10**	**.06**	.02	.00	**.12**

Note. *n* = 3,278-3,376. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

Table B.5. Descriptive Statistics and Reliability Estimates for Rational Biodata Inventory (RBI) Scale Scores

Scale	Items	*M*	*SD*	α
Peer Leadership	6	3.60	.65	.71
Cognitive Flexibility	8	3.48	.64	.76
Achievement	9	3.58	.57	.69
Fitness Motivation	7	3.31	.69	.74
Interpersonal Skills - Diplomacy	5	3.66	.75	.71
Stress Tolerance	11	2.99	.51	.67
Hostility to Authority	7	2.47	.64	.68
Self-Efficacy	6	4.02	.61	.77
Cultural Tolerance	5	3.75	.73	.69
Internal Locus of Control	8	3.55	.56	.66
Army Affective Commitment	7	3.71	.69	.71
Respect for Authority	4	3.54	.68	.66
Narcissism	6	3.60	.57	.55
Gratitude	3	3.98	.71	.42
Response Distortion Scale	7	0.09	.14	.52
Pure Fitness Motivation [a]	5	3.41	.73	.71

Note. *n* = 6,517-6,518. Items = number of items comprising each final scale. α = coefficient alpha. RBI scale scores range from 1 – 5, except for the Lie scale, which ranges from 0 – 1.

[a] An alternative version of the Fitness Motivation scale with the ability items removed.

Table B.6. Intercorrelations among RBI Scale Scores

	Scale	**1**	**2**	**3**	**4**	**5**	6	7	8	9	10	11	12	13	14	15
1	Peer Leadership															
2	Cognitive Flexibility	**.51**														
3	Achievement	**.55**	**.49**													
4	Fitness Motivation	**.29**	**.15**	**.28**												
5	Interpersonal Skills - Diplomacy	**.49**	**.29**	**.38**	**.22**											
6	Stress Tolerance	**.11**	**.13**	**.05**	**.20**	**.23**										
7	Hostility to Authority	**-.09**	**-.17**	**-.24**	**-.04**	**-.16**	**-.35**									
8	Self-Efficacy	**.56**	**.44**	**.57**	**.38**	**.46**	**.22**	**-.18**								
9	Cultural Tolerance	**.35**	**.42**	**.31**	**.12**	**.42**	**.30**	**-.34**	**.40**							
10	Internal Locus of Control	**.29**	**.27**	**.35**	**.20**	**.35**	**.41**	**-.38**	**.43**	**.37**						
11	Army Affective Commitment	**.30**	**.19**	**.29**	**.29**	**.27**	**.20**	**-.20**	**.42**	**.26**	**.32**					
12	Respect for Authority	**.27**	**.29**	**.49**	**.10**	**.20**	**-.03**	**-.19**	**.31**	**.20**	**.21**	**.19**				
13	Narcissism	**.37**	**.22**	**.35**	**.18**	**.22**	**-.16**	**.14**	**.40**	**.08**	**.09**	**.18**	**.17**			
14	Gratitude	**.27**	**.24**	**.34**	**.11**	**.32**	**.09**	**-.27**	**.35**	**.29**	**.34**	**.23**	**.34**	**.10**		
15	Response Distortion Scale	**.16**	**.15**	**.17**	**.12**	**.13**	**.23**	**-.20**	**.19**	**.20**	**.16**	**.12**	**.09**	**.03**	.00	
16	Pure Fitness Motivation [a]	**.32**	**.19**	**.32**	**.94**	**.24**	**.18**	**-.07**	**.41**	**.16**	**.22**	**.34**	**.14**	**.19**	**.16**	**.13**

Note. n = 6,516-6,518. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

[a] An alternative version of the Fitness Motivation scale with the ability items removed.

Table B.7. Descriptive Statistics and Reliability Estimates for Army Knowledge Assessment (AKA) Scales

Scale	Items	*M*	*SD*	α
Realistic Interests	5	4.06	.60	.76
Investigative Interests	5	3.37	.74	.83
Artistic Interests	5	2.72	.94	.89
Social Interests	5	3.79	.71	.83
Enterprising Interests	5	3.69	.71	.82
Conventional Interests	5	3.95	.69	.84

Note. n = 7,610-7,613. Items = number of items comprising each final scale. α = coefficient alpha. AKA scale scores range from 1 – 5.

Table B.8. Intercorrelations among AKA Scales

Scale		1	2	3	4	5
1	Realistic Interests					
2	Investigative Interests	.37				
3	Artistic Interests	.13	.51			
4	Social Interests	.38	.38	.29		
5	Enterprising Interests	.39	.37	.25	.47	
6	Conventional Interests	.43	.27	.07	.44	.51

Note. n = 7,585-7,613. All correlations are statistically significant, $p < .05$ (two-tailed).

Table B.9. Descriptive Statistics and Reliability Estimates for Work Preferences Assessment (WPA) Dimension and Facet Scores

Scale	Items	*M*	*SD*	α
Realistic Interests (D)	13	3.45	.81	.91
Mechanical (F)	5	3.13	1.07	.90
Physical (F)	7	3.70	.86	.89
Investigative Interests (D)	12	3.27	.65	.84
Critical Thinking (F)	6	3.77	.72	.82
Conduct Research (F)	6	2.78	.78	.76
Artistic Interests (D)	12	2.78	.77	.88
Artistic Activities (F)	8	2.37	.87	.85
Creativity (F)	4	3.59	.86	.82
Social Interests (D)	10	3.62	.65	.83
Work with Others (F)	5	3.83	.71	.77
Help Others (F)	5	3.42	.76	.72
Enterprising Interests (D)	13	3.36	.59	.81
Prestige (F)	5	3.88	.66	.67
Lead Others (F)	4	3.56	.74	.71
High Profile (F)	4	2.51	.88	.72
Conventional Interests (D)	12	3.23	.63	.82
Information Management (F)	6	2.64	.85	.82
Detail Orientation (F)	3	3.88	.78	.73
Clear Procedures (F)	3	3.91	.76	.65

Note. n = 7,511-7,512. D = Dimension. F = Facet. Items = number of items comprising each final scale.
α = coefficient alpha. WPA scale scores range from 1 – 5.

Table B.10. Intercorrelations among WPA Dimension and Facet Scores

Scale		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	Realistic Interests (D)																			
2	Mechanical (F)	**.83**																		
3	Physical (F)	**.86**	**.46**																	
4	Investigative Interests (D)	**.14**	**.12**	**.14**																
5	Critical Thinking (F)	**.17**	**.08**	**.22**	**.86**															
6	Conduct Research (F)	**.08**	**.13**	.02	**.88**	**.52**														
7	Artistic Interests (D)	**.11**	**.19**	.01	**.40**	**.23**	**.46**													
8	Artistic Activities (F)	**.09**	**.18**	-.02	**.31**	**.10**	**.42**	**.95**												
9	Creativity (F)	**.11**	**.13**	**.07**	**.46**	**.41**	**.39**	**.76**	**.51**											
10	Social Interests (D)	**.07**	**-.07**	**.17**	**.53**	**.52**	**.40**	**.28**	**.20**	**.33**										
11	Work with Others (F)	**.17**	-.01	**.29**	**.44**	**.49**	**.29**	**.19**	**.11**	**.29**	**.88**									
12	Help Others (F)	**-.05**	**-.11**	**.03**	**.50**	**.43**	**.43**	**.30**	**.25**	**.30**	**.90**	**.59**								
13	Enterprising Interests (D)	**.14**	**.05**	**.18**	**.59**	**.55**	**.48**	**.38**	**.29**	**.41**	**.58**	**.52**	**.51**							
14	Prestige (F)	**.15**	**.03**	**.22**	**.48**	**.54**	**.31**	**.18**	**.08**	**.33**	**.49**	**.48**	**.39**	**.80**						
15	Lead Others (F)	**.19**	.01	**.29**	**.47**	**.50**	**.32**	**.23**	**.14**	**.34**	**.58**	**.55**	**.49**	**.81**	**.57**					
16	High Profile (F)	.00	**.08**	**-.07**	**.45**	**.28**	**.50**	**.45**	**.45**	**.31**	**.31**	**.22**	**.33**	**.75**	**.33**	**.38**				
17	Conventional Interests (D)	**.10**	**.12**	**.06**	**.59**	**.51**	**.52**	**.25**	**.22**	**.22**	**.54**	**.46**	**.50**	**.58**	**.46**	**.41**	**.48**			
18	Information Management (F)	**-.05**	**.07**	**-.14**	**.49**	**.31**	**.53**	**.36**	**.36**	**.22**	**.40**	**.27**	**.43**	**.51**	**.28**	**.29**	**.61**	**.86**		
19	Detail Orientation (F)	**.21**	**.12**	**.25**	**.52**	**.59**	**.32**	**.06**	**-.02**	**.22**	**.47**	**.48**	**.36**	**.41**	**.47**	**.38**	**.12**	**.68**	**.30**	
20	Clear Procedures (F)	**.18**	**.09**	**.22**	**.46**	**.52**	**.29**	**.04**	**-.02**	**.16**	**.49**	**.49**	**.39**	**.40**	**.47**	**.37**	**.13**	**.73**	**.34**	**.89**

Note. n = 7,510-7,512. D = Dimension. F = Facet. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

APPENDIX C
SCALE-LEVEL CORRELATIONS BETWEEN SELECTED PREDICTOR AND CRITERION MEASURES

Table C.1. Correlations between Predictor Scale Scores and Selected Performance-Related Criterion Measures

	Criterion Measure/Scale										
Predictor Measure/Scale	MOS-SPEC JKT	MOS-SPEC PRS	EFFRT PRS	PHYS FIT PRS	APFT SCORE	PEERS PRS	LEAD PRS	PER DISC PRS	DISC INC	AIT SCORE	AIT PASS/ FAIL
AFQT	**.48**	**.14**	**.18**	**.08**	.04	**.15**	**.14**	**.17**	**-.09**	**.34**	**.04**
Assembling Objects (AO)	**.30**	**.16**	**.19**	**.12**	.01	**.13**	**.16**	**.14**	**-.11**	**.19**	**.07**
TAPAS-95s											
Achievement	.08	**.11**	.08	.08	**.12**	-.02	.05	.05	.01	**.18**	.02
Curiosity	**.11**	.01	-.02	-.01	.03	-.02	.01	-.01	-.04	**.09**	.02
Non-Delinquency	.05	.00	.03	-.04	-.07	.08	.01	**.16**	**-.11**	**.13**	.01
Dominance	.05	-.02	-.06	-.05	-.03	-.07	-.03	**-.12**	.03	.05	-.03
Even-Temper	**.14**	.07	.07	.04	-.05	.07	.02	**.15**	**-.09**	**.12**	**.09**
Attention-Seeking	.02	-.05	-.08	-.02	-.03	**-.09**	**-.09**	**-.17**	**.13**	-.05	-.03
Intellectual Efficiency	**.14**	.04	-.02	-.08	.00	-.05	-.05	-.04	.07	**.15**	.00
Order	.00	.05	-.01	-.02	.08	-.04	.00	.03	**-.10**	.05	-.02
Physical Conditioning	-.04	**.13**	.07	**.20**	**.24**	.02	**.17**	.00	-.02	-.03	**.08**
Tolerance	-.01	-.05	-.03	-.04	-.04	.04	.00	.00	.00	.04	.00
Cooperation/Trust	-.03	-.03	.04	.02	-.06	.04	-.02	**.11**	-.07	.01	.02
Optimism	**.15**	.06	.03	-.01	-.02	.02	.05	.02	.02	.08	**.08**
AIM											
Adjustment	**.14**	.08	**.13**	.05	.08	.10	**.14**	**.15**	.01	**.14**	**.10**
Agreeableness	.08	**.11**	**.11**	.03	.01	.09	**.11**	**.16**	-.08	**.10**	**.08**
Dependability	**.13**	**.14**	.09	.04	.10	.06	**.13**	**.16**	-.08	**.18**	**.05**
Leadership	**.17**	.02	.03	.07	**.18**	-.05	.10	.00	-.01	**.10**	.00
Physical Conditioning	.03	.09	**.10**	**.22**	**.25**	**.11**	**.21**	**.11**	-.05	.04	**.13**
Work Orientation	.07	.04	.03	.05	**.16**	-.06	.08	-.04	.06	**.08**	**.05**
RBI											
Peer Leadership	.04	.03	.00	.05	**.08**	.01	.03	-.06	.01	.04	.01
Cognitive Flexibility	.07	.01	.01	-.01	.04	.01	.01	-.01	.02	**.09**	**.04**
Achievement	-.06	.01	-.02	.06	**.12**	-.03	.03	**-.08**	.05	.05	**.04**
Fitness Motivation	**.09**	**.13**	**.11**	**.26**	**.34**	**.07**	**.17**	.07	-.05	.01	**.15**
Interpersonal Skills - Diplomacy	-.02	.00	-.02	.04	**.08**	.02	.05	-.06	**.08**	.06	**.06**
Stress Tolerance	**.13**	**.10**	.05	.06	.07	.03	.07	.03	-.04	**.09**	**.10**
Hostility to Authority	**-.15**	-.02	-.04	.00	-.06	-.04	.01	-.05	.07	**-.17**	**-.04**

Table C.1. Correlations between Predictor Scale Scores and Selected Performance-Related Criterion Measures (cont'd)

	Criterion Measure/Scale										
Predictor Measure/Scale	MOS-SPEC JKT	MOS-SPEC PRS	EFFRT PRS	PHYS FIT PRS	APFT SCORE	PEERS PRS	LEAD PRS	PER DISC PRS	DISC INC	AIT SCORE	AIT PASS/ FAIL
RBI (continued)											
Self-Efficacy	.04	.03	-.01	.05	.04	-.02	.02	-.05	.02	.07	**.08**
Cultural Tolerance	**.07**	.07	.00	.02	.00	**.08**	.02	.02	.01	**.09**	.02
Internal Locus of Control	**.14**	**.09**	**.10**	**.14**	.07	**.11**	**.11**	**.09**	-.02	**.16**	**.07**
Army Affective Commitment	**.11**	.02	.03	.06	.02	.02	.05	.04	-.05	.05	**.12**
Respect for Authority	.00	.00	.03	.03	.01	.04	.04	.02	.05	.03	**.05**
Narcissism	**-.08**	-.02	-.07	.03	.02	**-.08**	-.02	**-.10**	**.09**	-.04	-.01
Gratitude	**.09**	.01	.04	.04	-.01	.03	.02	.03	-.03	**.10**	**.04**
PSJT	**.21**	**.08**	**.10**	.03	.03	.06	.07	**.10**	**-.07**	**.31**	**.05**
AKA											
Realistic Interests	**.10**	.03	.03	.04	.00	.05	.01	.02	.04	**.07**	**.05**
Investigative Interests	**-.11**	.03	-.04	.00	-.01	-.04	-.03	-.05	.05	-.03	.00
Artistic Interests	**-.19**	-.05	**-.09**	-.05	.01	**-.09**	-.04	**-.08**	**.07**	**-.11**	.01
Social Interests	.03	.00	-.01	.00	-.01	.01	-.01	-.03	.06	.05	**.03**
Enterprising Interests	.05	.03	.03	.02	.02	.04	.01	.00	.03	.05	.01
Conventional Interests	**.13**	.00	.04	.00	-.01	.03	.00	.00	.03	**.12**	**.05**
WPA											
Realistic Interests	-.02	.04	.00	.05	.02	-.04	.03	.00	.01	-.06	**.11**
Mechanical	-.03	-.01	-.02	.00	-.03	**-.07**	.03	-.03	.01	**-.07**	**.06**
Physical	.00	**.08**	.03	**.10**	**.08**	.02	.02	.03	.01	-.04	**.11**
Investigative Interests	**-.07**	.02	-.01	.00	.01	-.02	-.01	-.03	.05	.03	.03
Critical Thinking	-.01	**.06**	.03	.04	.03	.01	.02	.00	.05	**.08**	**.04**
Conduct Research	**-.11**	-.04	-.05	-.04	-.01	-.05	-.04	**-.06**	.03	-.02	.01
Artistic Interests	**-.16**	**-.07**	**-.06**	**-.07**	-.03	-.04	-.06	-.05	.02	-.03	**-.05**
Artistic Activities	**-.17**	**-.08**	**-.09**	**-.08**	-.03	-.05	**-.07**	**-.07**	.02	-.06	**-.06**
Creativity	**-.08**	-.02	.01	-.03	.00	.00	.00	.00	.00	.05	-.02
Social Interests	**-.16**	.02	-.02	.01	.02	.01	-.03	-.03	.05	-.02	.00
Work with Others	**-.14**	.04	-.02	.02	.00	.02	-.03	-.01	**.06**	-.03	**.04**
Help Others	**-.14**	.00	-.02	.00	.04	-.01	-.02	-.04	.03	-.01	**-.03**

Table C.1. Correlations between Predictor Scale Scores and Selected Performance-Related Criterion Measures (cont'd)

	Criterion Measure/Scale										
Predictor Measure/Scale	MOS-SPEC JKT	MOS-SPEC PRS	EFFRT PRS	PHYS FIT PRS	APFT SCORE	PEERS PRS	LEAD PRS	PER DISC PRS	DISC INC	AIT SCORE	AIT PASS/ FAIL
WPA (continued)											
Enterprising Interests	**-.10**	.00	-.04	.02	.01	-.04	.00	**-.07**	**.08**	.01	**.04**
Prestige	.01	.04	.00	.04	-.01	.00	.01	-.01	**.08**	**.08**	**.05**
Lead Others	**-.06**	.03	-.02	.03	.05	-.01	.01	**-.07**	**.08**	.02	**.04**
High Profile	**-.16**	-.05	**-.07**	-.03	.00	**-.09**	-.01	**-.08**	.03	**-.07**	.00
Conventional Interests	**-.16**	-.03	**-.07**	-.01	-.02	-.05	-.03	-.05	**.09**	-.02	.01
Information Management	**-.19**	**-.09**	**-.09**	-.05	-.03	**-.08**	-.04	**-.08**	**.08**	-.04	-.01
Detail Orientation	-.06	.03	-.01	.02	.03	.01	.01	-.01	.05	.02	.02
Clear Procedures	**-.07**	.02	-.02	.02	.00	.00	.00	.00	.06	.02	.02

Note. AFQT *n* = 1,101 – 5,227. AO *n* = 973 – 4,685. TAPAS-95s *n* = 505 – 2,535. AIM *n* = 356 – 2,445. RBI *n* = 797 – 4,061. PSJT *n* = 293 – 2,390. AKA *n* = 1,009 – 4,931. WPA *n* = 981 – 4,884. Statistically significant correlations are bolded, $p < .05$ (two-tailed). MOS-SPEC JKT = MOS-Specific Job Knowledge Test, MOS-SPEC PRS = MOS-Specific Performance Rating Scale (PRS) Composite, EFFRT PRS = Effort PRS Composite, PHYS FIT PRS = Physical Fitness & Military Bearing PRS Composite, APFT SCORE = Army Physical Fitness Test (APFT) Score, PEERS PRS = Support for Peers PRS Composite, LEAD PRS = Peer Leadership PRS Composite, PER DISC PRS = Personal Discipline PRS Composite, DISC INC = Disciplinary Incidence (0 = *None*, 1 = *One or more*), AIT SCORE = Standardized Average AIT/OSUT Exam Grade, AIT PASS/FAIL = Graduation from AIT/OSUT (0 = *Discharged from Army*, 1 = *Graduated from AIT/OSUT*).

Table C.2. Correlations between Predictor Scale Scores and Selected Retention-Related Criterion Measures

	Criterion Measure/Scale					
Predictor Measure/Scale	AFFECT COMMIT	ARMY FIT	CAR INTENT	ATTRIT COG	6-MO ATTRIT	ARMY ADJUST
AFQT	**-.06**	.01	-.01	**.07**	-.02	**.13**
Assembling Objects (AO)	-.01	.04	-.04	.06	**-.07**	**.15**
TAPAS-95s						
Achievement	**.16**	**.19**	**.12**	**.16**	-.04	**.24**
Curiosity	.04	.04	**.13**	.05	-.04	.07
Non-Delinquency	.03	.06	.00	.08	.01	.05
Dominance	**.09**	.09	.08	.06	.03	.07
Even-Temper	.04	.07	.08	.08	**-.10**	.08
Attention-Seeking	.03	.02	.03	-.02	.04	.01
Intellectual Efficiency	.07	.07	**.09**	**.12**	-.03	.04
Order	.00	-.02	-.04	-.03	-.01	.04
Physical Conditioning	.07	**.14**	.07	**.09**	**-.10**	**.19**
Tolerance	.06	.05	**.15**	.07	-.02	**.11**
Cooperation/Trust	**-.10**	-.08	**-.14**	-.07	-.02	-.06
Optimism	.07	**.10**	**.13**	**.09**	**-.10**	**.14**
AIM						
Adjustment	**.18**	**.23**	**.17**	**.18**	**-.12**	**.22**
Agreeableness	.09	**.13**	.04	.08	**-.10**	.09
Dependability	**.20**	**.21**	.09	**.18**	**-.08**	**.17**
Leadership	**.20**	**.25**	**.20**	**.17**	-.04	**.22**
Physical Conditioning	**.19**	**.23**	**.12**	**.19**	**-.12**	**.24**
Work Orientation	**.19**	**.20**	**.17**	**.11**	**-.10**	**.24**
RBI						
Peer Leadership	**.22**	**.22**	**.16**	**.18**	-.03	**.17**
Cognitive Flexibility	**.14**	**.17**	**.12**	**.13**	**-.06**	**.14**
Achievement	**.26**	**.25**	**.11**	**.18**	-.03	**.16**
Fitness Motivation	**.13**	**.20**	**.12**	**.16**	**-.11**	**.27**
Interpersonal Skills - Diplomacy	**.13**	**.17**	**.07**	**.13**	**-.07**	**.15**
Stress Tolerance	.01	**.12**	.04	**.14**	**-.07**	**.24**
Hostility to Authority	-.03	**-.12**	.03	**-.10**	**.04**	**-.15**
Self-Efficacy	**.22**	**.23**	**.14**	**.20**	**-.10**	**.21**
Cultural Tolerance	**.09**	**.15**	.01	**.10**	**-.04**	**.16**
Internal Locus of Control	**.14**	**.23**	.04	**.19**	**-.06**	**.21**
Army Affective Commitment	**.37**	**.31**	**.25**	**.25**	**-.13**	**.25**
Respect for Authority	**.22**	**.20**	**.09**	**.14**	-.03	**.11**
Narcissism	**.17**	**.14**	**.09**	**.07**	.00	.03
Gratitude	**.10**	**.10**	-.04	**.08**	**-.06**	**.08**
PSJT	-.03	.05	.01	.05	**-.06**	**.14**

Table C.2. Correlations between Predictor Scale Scores and Selected Retention-Related Criterion Measures (cont'd)

	Criterion Measure/Scale					
Predictor Measure/Scale	AFFECT COMMIT	ARMY FIT	CAR INTENT	ATTRIT COG	6-MO ATTRIT	ARMY ADJUST
AKA						
Realistic Interests	**.16**	**.16**	**.11**	**.15**	**-.04**	**.11**
Investigative Interests	**.10**	**.06**	**.07**	.03	-.01	.04
Artistic Interests	**.09**	.03	**.09**	.01	-.02	-.03
Social Interests	**.07**	**.07**	**.06**	.05	-.02	.04
Enterprising Interests	**.10**	**.08**	**.07**	**.07**	-.01	**.07**
Conventional Interests	**.10**	**.09**	**.07**	**.11**	**-.05**	**.06**
WPA						
Realistic Interests	**.19**	**.15**	**.17**	**.11**	**-.07**	**.12**
Mechanical	**.08**	.01	**.08**	.00	**-.04**	.01
Physical	**.21**	**.24**	**.19**	**.17**	**-.07**	**.19**
Investigative Interests	**.11**	**.11**	**.14**	**.09**	**-.04**	**.06**
Critical Thinking	**.13**	**.14**	**.15**	**.14**	**-.04**	**.11**
Conduct Research	.06	.05	**.09**	.01	-.03	.00
Artistic Interests	.00	-.03	.02	-.03	.01	**-.10**
Artistic Activities	-.02	-.06	.01	**-.06**	.02	**-.13**
Creativity	.05	.04	.04	.04	-.01	-.03
Social Interests	**.15**	**.17**	**.10**	**.10**	-.02	**.07**
Work with Others	**.16**	**.17**	**.11**	**.10**	**-.05**	**.08**
Help Others	**.11**	**.14**	**.08**	**.08**	.02	.05
Enterprising Interests	**.17**	**.17**	**.14**	**.08**	**-.04**	**.08**
Prestige	**.14**	**.16**	**.08**	**.08**	**-.05**	**.07**
Lead Others	**.22**	**.22**	**.19**	**.15**	-.03	**.16**
High Profile	.04	.02	.06	-.02	-.01	-.02
Conventional Interests	**.13**	**.13**	**.10**	.06	**-.04**	.05
Information Management	.05	.04	.05	-.02	-.02	-.02
Detail Orientation	**.17**	**.18**	**.14**	**.16**	**-.04**	**.12**
Clear Procedures	**.15**	**.17**	**.10**	**.13**	**-.04**	**.11**

Note. AFQT n = 1,136 – 3,203. AO n = 1001 – 2,960. TAPAS-95s n = 516 – 1,696. AIM n = 353 – 1,700. RBI n = 819 – 2,488. PSJT n = 732 – 1,253. AKA n = 1,088 – 3,003. WPA n = 1,088 – 2,968. Statistically significant correlations are bolded, $p < .05$ (two-tailed). AFFECT COMMIT = Army Life Questionnaire (ALQ) Army Affective Commitment scale, ARMY FIT = ALQ Needs-Supplies Army Fit scale, CAR INTENT = ALQ Career Intentions scale, ATTRIT COG = ALQ Attrition Cognitions scale, 6-MO ATTRIT = 6-Month Attrition, ARMY ADJUST = ALQ Adjustment to Army Life scale.

Table C.3. Correlations between the AFQT and Scale Scores from the Experimental Predictor Measures

Predictor Measure/Scale	*n*	AFQT
AO	7,300	**.42**
TAPAS-95s		
Achievement	3,362	**.06**
Curiosity	3,362	**.24**
Non-Delinquency	3,362	**.06**
Dominance	3,362	**.06**
Even-Temper	3,362	**.14**
Attention-Seeking	3,362	**-.07**
Intellectual Efficiency	3,362	**.38**
Order	3,362	**-.04**
Physical Conditioning	3,362	.00
Tolerance	3,362	.02
Cooperation/Trust	3,362	**-.04**
Optimism	3,362	**.18**
AIM		
Adjustment	3,343	**.12**
Agreeableness	3,262	**.11**
Dependability	3,313	**.12**
Leadership	3,351	**.14**
Physical Conditioning	3,307	**.04**
Work Orientation	3,302	.02
RBI		
Peer Leadership	6,482	**.13**
Cognitive Flexibility	6,482	**.26**
Achievement	6,482	**.05**
Fitness Motivation	6,482	**.04**
Interpersonal Skills - Diplomacy	6,482	**.03**
Stress Tolerance	6,482	**.17**
Hostility to Authority	6,482	**-.18**
Self-Efficacy	6,482	**.06**
Cultural Tolerance	6,482	**.09**
Internal Locus of Control	6,481	**.19**
Army Affective Commitment	6,482	**.03**
Respect for Authority	6,481	**-.05**
Narcissism	6,482	**-.07**
Gratitude	6,482	**.13**
PSJT	3,981	**.25**
AKA		
Realistic Interests	7,567	**.07**
Investigative Interests	7,565	**-.18**
Artistic Interests	7,567	**-.32**
Social Interests	7,568	-.01
Enterprising Interests	7,568	.02
Conventional Interests	7,543	**.15**

Table C.3. (Continued)

Predictor Measure/Scale	*n*	AFQT
WPA		
Realistic Interests	7,466	**-.12**
Mechanical	7,466	**-.10**
Physical	7,466	**-.10**
Investigative Interests	7,466	**.08**
Critical Thinking	7,465	**.15**
Conduct Research	7,466	-.01
Artistic Interests	7,466	**-.07**
Artistic Activities	7,465	**-.11**
Creativity	7,465	**.04**
Social Interests	7,466	**-.13**
Work with Others	7,466	**-.15**
Help Others	7,466	**-.09**
Enterprising Interests	7,466	**-.05**
Prestige	7,466	**.04**
Lead Others	7,465	**-.06**
High Profile	7,466	**-.09**
Conventional Interests	7,466	**-.21**
Information Management	7,466	**-.19**
Detail Orientation	7,466	**-.09**
Clear Procedures	7,466	**-.14**

Note. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

Table C.4. Correlations between Scales Scores from the TAPAS-95s and Other Temperament Predictor Measures

	TAPAS-95s Scale											
Measure/Scale	ACH	CUR	DEL	DOM	TEM	ATT	INT	ORD	PHY	TOL	TRU	OPT
AIM												
Adjustment	**.11**	**.19**	**.14**	**.05**	**.33**	**-.18**	**.13**	-.01	**.08**	**.12**	-.03	**.39**
Agreeableness	**.06**	**.14**	**.25**	**-.05**	**.43**	**-.26**	**.05**	-.01	**.05**	**.06**	**.08**	**.19**
Dependability	**.14**	**.16**	**.46**	**.10**	**.16**	**-.32**	**.07**	**.10**	-.04	**.04**	-.01	**.04**
Leadership	**.18**	**.21**	.02	**.51**	.01	**.06**	**.24**	**.04**	**.08**	**.12**	**-.25**	**.05**
Physical Conditioning	**.20**	**.07**	-.01	**.04**	**.06**	**-.06**	.03	**.05**	**.62**	**.04**	**-.14**	.03
Work Orientation	**.34**	**.22**	.03	**.21**	**.10**	**-.07**	**.18**	**.09**	**.30**	**.13**	**-.25**	**.06**
RBI												
Peer Leadership	**.14**	**.21**	.00	**.43**	.03	**.09**	**.25**	.02	**.13**	**.17**	**-.22**	**.06**
Cognitive Flexibility	**.12**	**.42**	**.05**	**.17**	**.17**	**-.10**	**.36**	-.02	.02	**.24**	**-.12**	**.07**
Achievement	**.23**	**.20**	**.16**	**.24**	.00	**-.08**	**.15**	**.12**	**.12**	**.13**	**-.16**	-.04
Fitness Motivation	**.15**	**.05**	**-.12**	**.08**	.04	.02	**.06**	-.03	**.62**	.01	**-.18**	**.07**
Interpersonal Skills - Diplomacy	**.07**	**.12**	-.01	**.32**	**.05**	**.18**	**.10**	.00	**.10**	**.16**	**-.08**	**.11**
Stress Tolerance	**.14**	**.17**	**.04**	**.07**	**.25**	**-.11**	**.20**	-.01	**.14**	**.09**	**-.05**	**.31**
Hostility to Authority	**-.13**	**-.18**	**-.44**	**-.06**	**-.18**	**.36**	**-.09**	**-.10**	**.06**	**-.05**	**-.07**	**-.08**
Self-Efficacy	**.24**	**.20**	.03	**.26**	**.11**	-.04	**.21**	**.05**	**.19**	**.15**	**-.19**	**.15**
Cultural Tolerance	**.11**	**.22**	**.15**	**.18**	**.17**	**-.11**	**.15**	.00	-.01	**.34**	-.02	**.12**
Internal Locus of Control	**.18**	**.17**	**.11**	**.14**	**.15**	**-.09**	**.17**	**.09**	**.10**	**.10**	**-.05**	**.21**
Army Affective Commitment	**.18**	**.09**	**.09**	**.13**	**.09**	**-.08**	.03	.02	**.12**	**.10**	**-.12**	**.13**
Respect for Authority	**.13**	**.07**	**.17**	**.08**	.01	**-.07**	-.01	.04	.01	**.07**	**-.06**	**-.05**
Narcissism	**.07**	**.07**	**-.08**	**.21**	**-.11**	**.10**	**.10**	**.08**	**.11**	**.09**	**-.17**	.00
Gratitude	**.11**	**.12**	**.17**	**.11**	**.08**	**-.07**	**.05**	**.04**	.01	**.06**	.01	**.07**
PSJT	**.14**	**.17**	**.38**	**.12**	**.18**	**-.26**	.05	.05	-.05	**.13**	.09	**.10**

Note. AIM n = 2,618 – 2,650. RBI n = 2,446. PSJT n = 450. Statistically significant correlations are bolded, $p < .05$ (two-tailed). ACH = Achievement, CUR = Curiosity, DEL = Non-Delinquency, DOM = Dominance, TEM = Even-Temper, ATT = Attention-Seeking, INT = Intellectual Efficiency, ORD = Order, PHY = Physical Conditioning TOL = Tolerance, TRU = Cooperation/Trust, OPT = Optimism.

Table C.5. Correlations between Scale Scores from the WPA and the AKA

	AKA Scale					
WPA Scale	REAL	INVEST	ART	SOC	ENTER	CONV
Realistic Interests	**.14**	**.12**	**.14**	**.10**	**.07**	**.04**
Mechanical	**.07**	**.10**	**.14**	**.05**	**.03**	-.01
Physical	**.17**	**.11**	**.10**	**.12**	**.09**	**.07**
Investigative Interests	**.20**	**.20**	**.12**	**.20**	**.20**	**.20**
Critical Thinking	**.26**	**.16**	**.05**	**.21**	**.22**	**.24**
Conduct Research	**.10**	**.18**	**.16**	**.14**	**.13**	**.12**
Artistic Interests	**.05**	**.16**	**.18**	**.10**	**.10**	**.06**
Artistic Activities	.00	**.15**	**.19**	**.06**	**.06**	.02
Creativity	**.13**	**.13**	**.10**	**.13**	**.14**	**.12**
Social Interests	**.22**	**.22**	**.16**	**.26**	**.20**	**.18**
Work with Others	**.23**	**.20**	**.14**	**.25**	**.19**	**.18**
Help Others	**.17**	**.19**	**.15**	**.22**	**.17**	**.15**
Enterprising Interests	**.20**	**.20**	**.14**	**.19**	**.19**	**.17**
Prestige	**.24**	**.15**	**.05**	**.19**	**.19**	**.21**
Lead Others	**.20**	**.17**	**.11**	**.18**	**.17**	**.16**
High Profile	**.03**	**.15**	**.17**	**.08**	**.09**	**.05**
Conventional Interests	**.17**	**.26**	**.25**	**.21**	**.19**	**.15**
Information Management	**.05**	**.21**	**.24**	**.13**	**.12**	**.07**
Detail Orientation	**.24**	**.19**	**.13**	**.21**	**.19**	**.18**
Clear Procedures	**.24**	**.21**	**.15**	**.21**	**.20**	**.18**

Note. n = 7,279 – 7,302. Statistically significant correlations are bolded, $p < .05$ (two-tailed). REAL = Realistic Interests, INVEST = Investigative Interests, ART = Artistic Interests, SOC = Social Interests, ENTER = Enterprising Interests, CONV = Conventional Interests.

Table C.6. Correlations between Scale Scores from the TAPAS-95s and the WPA

	TAPAS-95s Scale											
WPA Scale	ACH	CUR	DEL	DOM	TEM	ATT	INT	ORD	PHY	TOL	TRU	OPT
Realistic Interests	**.11**	-.02	**-.10**	**-.05**	.01	.02	**-.08**	**-.04**	**.23**	**-.06**	**-.08**	**.04**
Mechanical	**.08**	.02	**-.11**	**-.11**	.01	.00	**-.05**	-.03	**.07**	**-.08**	**-.04**	.02
Physical	**.11**	-.03	**-.07**	.01	.02	**.04**	**-.07**	**-.04**	**.34**	-.01	**-.09**	**.04**
Investigative Interests	**.16**	**.40**	**.08**	**.13**	**.14**	**-.13**	**.27**	**.05**	.01	**.17**	**-.10**	**.03**
Critical Thinking	**.20**	**.34**	**.09**	**.18**	**.16**	**-.12**	**.28**	**.05**	**.06**	**.15**	**-.12**	**.08**
Conduct Research	**.08**	**.36**	**.06**	**.05**	**.09**	**-.10**	**.19**	.03	-.03	**.15**	**-.05**	-.01
Artistic Interests	**-.05**	**.17**	**-.09**	.02	**.05**	.01	.03	-.02	**-.04**	**.13**	.01	**-.06**
Artistic Activities	**-.09**	**.11**	**-.08**	**-.04**	.02	.02	-.03	-.03	**-.04**	**.11**	.03	**-.08**
Creativity	**.04**	**.24**	**-.07**	**.12**	**.10**	-.01	**.15**	-.01	-.02	**.13**	-.03	.01
Social Interests	**.04**	**.10**	**.13**	**.20**	**.07**	-.03	**-.04**	.03	.02	**.16**	-.01	**-.07**
Work with Others	**.04**	**.08**	**.09**	**.16**	**.09**	.00	**-.05**	.02	**.06**	**.14**	.00	-.03
Help Others	**.04**	**.10**	**.14**	**.20**	**.04**	**-.06**	-.02	**.04**	-.02	**.15**	-.03	**-.10**
Enterprising Interests	**.09**	**.13**	-.02	**.29**	.02	**.09**	**.06**	**.06**	**.08**	**.12**	**-.14**	**-.07**
Prestige	**.12**	**.13**	**.07**	**.23**	**.04**	.03	**.08**	**.10**	**.07**	**.10**	**-.10**	-.03
Lead Others	**.09**	**.09**	**-.05**	**.36**	.01	**.12**	**.05**	.02	**.10**	**.12**	**-.15**	-.03
High Profile	.00	**.08**	**-.07**	**.12**	-.01	**.08**	.02	.02	.01	**.08**	**-.08**	**-.09**
Conventional Interests	**.11**	.08	**.16**	**.09**	.02	**-.09**	-.02	**.18**	-.03	**.07**	**-.04**	**-.08**
Information Management	**.03**	**.06**	**.06**	**.06**	-.01	**-.04**	-.02	**.10**	**-.06**	**.07**	-.02	**-.10**
Detail Orientation	**.19**	**.15**	**.14**	**.10**	**.07**	**-.12**	**.09**	**.18**	**.04**	**.07**	**-.07**	.00
Clear Procedures	**.15**	**.09**	**.19**	**.08**	**.04**	**-.13**	.01	**.20**	.01	**.06**	**-.05**	-.02

Note. n = 3,183 – 3,184. Statistically significant correlations are bolded, $p < .05$ (two-tailed). ACH = Achievement, CUR = Curiosity, DEL = Non-Delinquency, DOM = Dominance, TEM = Even-Temper, ATT = Attention-Seeking, INT = Intellectual Efficiency, ORD = Order, PHY = Physical Conditioning TOL = Tolerance, TRU = Cooperation/Trust, OPT = Optimism. Define D and F

Table C.7. Intercorrelations among Scale Scores from Selected Performance-Related Criterion Measures

Scale		1	2	3	4	5	6	7	8	9	10
1	MOS-Specific Job Knowledge Test										
2	MOS-Specific PRS Composite	**.23**									
3	Effort PRS Composite	**.29**	**.63**								
4	Physical Fitness & Bearing PRS Composite	**.18**	**.57**	**.72**							
5	APFT Score	.02	**.18**	**.23**	**.44**						
6	Support for Peers PRS Composite	**.20**	**.57**	**.72**	**.60**	**.12**					
7	Peer Leadership PRS Composite	**.19**	**.60**	**.75**	**.68**	**.26**	**.69**				
8	Personal Discipline PRS Composite	**.24**	**.55**	**.78**	**.61**	**.12**	**.78**	**.65**			
9	Disciplinary Incidents	**-.12**	**-.19**	**-.26**	**-.18**	**-.11**	**-.16**	**-.21**	**-.27**		
10	Average AIT Exam Grade	--[a]	--	--	--	--	--	--	--	--	
11	Graduation from AIT/OSUT	.00	.03	.05	.04	-.02	.03	**.06**	.00	.00	**.41**

Note. n = 576 – 1,187. APFT = Army Physical Fitness Test. Disciplinary Incidents is a constructed variable based on the self-reported number of disciplinary incidents and is coded 0 = *None* and 1 = *One or more*. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

[a] Sample size too small to compute a meaningful estimate (n = 1 or 2).

Table C.8. Intercorrelations among Scale Scores from Selected Retention-Related Criterion Measures

Scale		1	2	3	4	5
1	Affective Commitment					
2	Army Fit	**.79**				
3	Career Intentions	**.56**	**.54**			
4	Attrition Cognitions	**.65**	**.71**	**.54**		
5	6-Month Attrition	**-.11**	**-.09**	**-.09**	**-.16**	
6	Adjustment to Army Life	**.43**	**.63**	**.38**	**.51**	-.06

Note. n = 679 – 1,133. Statistically significant correlations are bolded, $p < .05$ (two-tailed).

APPENDIX D
PREDICTOR SCORE SUBGROUP DIFFERENCES

Table D.1. Standardized Mean Differences (Cohen's d) by Subgroup Combination and Predictor Measure

	Gender Differences					Race Differences					Ethnicity Differences				
	Female		Male		F-M	Black (B)		White (W)		B-W	Hispanic (H)		White, Non-Hispanic (WNH)		H-WNH
Predictor	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
AFQT	53.97	18.52	58.21	20.49	-0.23	47.37	17.00	59.55	20.09	-0.72	51.62	18.43	60.81	20.02	-0.50
AO	53.89	7.81	55.49	7.96	-0.20	51.07	8.61	55.94	7.53	-0.57	54.88	7.72	56.06	7.53	-0.15
AIM															
Adjustment	1.23	0.32	1.26	0.28	-0.09	1.28	0.26	1.25	0.30	0.12	1.29	0.27	1.24	0.30	0.19
Agreeableness	1.27	0.29	1.26	0.26	0.03	1.29	0.26	1.25	0.27	0.15	1.29	0.25	1.24	0.27	0.20
Dependability	1.35	0.27	1.27	0.28	0.30	1.33	0.26	1.27	0.29	0.23	1.30	0.27	1.27	0.29	0.11
Leadership	1.25	0.28	1.20	0.28	0.18	1.25	0.25	1.20	0.28	0.20	1.21	0.27	1.20	0.29	0.04
Physical Conditioning	1.14	0.34	1.21	0.34	-0.21	1.22	0.30	1.20	0.35	0.07	1.22	0.31	1.19	0.35	0.10
Work Orientation	1.21	0.28	1.18	0.28	0.11	1.25	0.25	1.17	0.29	0.32	1.20	0.26	1.17	0.29	0.12
Average Absolute d					*0.15*					*0.18*					*0.12*
TAPAS															
Achievement	0.26	0.63	0.14	0.62	0.19	0.13	0.58	0.17	0.64	-0.07	0.15	0.63	0.17	0.64	-0.03
Curiosity	-0.03	0.82	-0.10	0.79	0.09	-0.02	0.75	-0.11	0.81	0.12	-0.06	0.81	-0.11	0.81	0.06
Non-Delinquency	0.30	0.66	0.08	0.66	0.33	0.11	0.62	0.12	0.67	-0.02	0.06	0.62	0.13	0.67	-0.11
Dominance	-0.03	0.60	-0.17	0.60	0.23	-0.03	0.58	-0.16	0.60	0.22	-0.17	0.60	-0.15	0.61	-0.03
Even-Temper	-0.56	0.82	-0.47	0.75	-0.11	-0.46	0.73	-0.50	0.77	0.05	-0.50	0.75	-0.50	0.77	0.00
Attention-Seeking	-0.15	0.79	-0.12	0.80	-0.04	-0.16	0.81	-0.11	0.80	-0.06	-0.13	0.81	-0.11	0.79	-0.02
Intellectual Efficiency	-0.27	0.64	-0.16	0.65	-0.17	-0.15	0.60	-0.19	0.66	0.07	-0.25	0.61	-0.18	0.66	-0.11
Order	0.11	0.62	-0.07	0.64	0.29	0.05	0.58	-0.05	0.65	0.17	-0.02	0.63	-0.05	0.65	0.05
Physical Conditioning	-0.04	0.72	0.17	0.70	-0.29	0.15	0.68	0.13	0.72	0.03	0.15	0.72	0.13	0.72	0.03
Tolerance	-0.28	0.62	-0.45	0.68	0.27	-0.24	0.67	-0.46	0.67	0.33	-0.34	0.62	-0.48	0.68	0.23
Cooperation/Trust	-0.20	0.85	-0.30	0.87	0.12	-0.30	0.88	-0.28	0.86	-0.02	-0.29	0.84	-0.28	0.87	-0.01
Optimism	-0.13	0.61	-0.06	0.60	-0.11	-0.08	0.59	-0.07	0.61	-0.02	-0.07	0.64	-0.07	0.60	0.00
Average Absolute d					*0.19*					*0.10*					*0.06*
PSJT	4.79	0.36	4.66	0.41	0.36	4.62	0.44	4.70	0.39	-0.18	4.67	0.38	4.71	0.40	-0.11

Table D.1. Standardized Mean Differences (Cohen's d) by Subgroup Combination and Predictor Measure (cont'd)

	Gender Differences					Race Differences					Ethnicity Differences				
	Female		Male		F-M	Black (B)		White (W)		B-W	Hispanic (H)		White, Non-Hispanic (WNH)		H-WNH
Predictor	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
RBI															
Peer Leadership	3.71	0.64	3.57	0.64	0.22	3.68	0.67	3.59	0.64	0.13	3.58	0.64	3.59	0.64	-0.02
Cognitive Flexibility	3.56	0.62	3.45	0.64	0.18	3.55	0.62	3.46	0.65	0.15	3.53	0.63	3.45	0.65	0.13
Achievement	3.76	0.56	3.53	0.56	0.41	3.77	0.58	3.54	0.56	0.40	3.60	0.58	3.53	0.56	0.12
Fitness Motivation	2.93	0.65	3.42	0.66	-0.75	3.27	0.72	3.32	0.68	-0.07	3.30	0.67	3.32	0.69	-0.03
Interpersonal Skills - Diplomacy	3.83	0.74	3.61	0.74	0.30	3.79	0.71	3.64	0.75	0.21	3.69	0.74	3.64	0.75	0.07
Stress Tolerance	2.92	0.54	3.01	0.50	-0.17	3.02	0.53	2.99	0.50	0.06	3.00	0.51	2.99	0.50	0.02
Hostility to Authority	2.23	0.59	2.54	0.64	-0.53	2.48	0.67	2.47	0.64	0.01	2.48	0.66	2.47	0.64	0.02
Self-Efficacy	4.11	0.58	4.00	0.62	0.19	4.18	0.59	4.00	0.61	0.31	4.04	0.61	3.99	0.61	0.08
Cultural Tolerance	3.95	0.64	3.69	0.74	0.41	3.88	0.69	3.72	0.73	0.23	3.98	0.70	3.68	0.73	0.43
Internal Locus of Control	3.65	0.54	3.52	0.57	0.24	3.58	0.55	3.55	0.57	0.05	3.54	0.56	3.55	0.57	-0.02
Army Affective Commitment	3.66	0.70	3.72	0.68	-0.09	3.57	0.68	3.74	0.68	-0.25	3.73	0.67	3.74	0.68	-0.01
Respect for Authority	3.67	0.67	3.50	0.68	0.25	3.63	0.74	3.52	0.68	0.15	3.56	0.68	3.51	0.67	0.07
Narcissism	3.63	0.56	3.60	0.57	0.05	3.83	0.59	3.56	0.55	0.46	3.66	0.57	3.55	0.55	0.19
Gratitude	4.13	0.65	3.94	0.72	0.29	3.89	0.75	4.00	0.70	-0.15	3.96	0.73	4.01	0.69	-0.07
Average Absolute d					*0.29*					*0.19*					*0.09*
AKA															
Realistic	4.09	0.58	4.05	0.61	0.07	4.08	0.63	4.06	0.59	0.03	4.04	0.62	4.06	0.59	-0.03
Investigative	3.42	0.74	3.36	0.74	0.08	3.51	0.73	3.34	0.74	0.23	3.41	0.75	3.33	0.74	0.11
Artistic	2.70	0.96	2.72	0.93	-0.02	2.97	0.94	2.66	0.93	0.33	2.81	0.93	2.64	0.93	0.18
Social	3.87	0.68	3.76	0.72	0.16	3.86	0.74	3.77	0.70	0.12	3.79	0.73	3.77	0.70	0.03
Enterprising	3.79	0.70	3.67	0.72	0.17	3.77	0.74	3.68	0.71	0.12	3.69	0.70	3.68	0.71	0.01
Conventional	4.02	0.67	3.93	0.70	0.13	3.95	0.71	3.95	0.69	0.00	3.90	0.69	3.96	0.69	-0.09
Average Absolute d					*0.11*					*0.14*					*0.08*

Table D.1. Standardized Mean Differences (Cohen's d) by Subgroup Combination and Predictor Measure (cont'd)

Predictor	Gender Differences					Race Differences					Ethnicity Differences				
	Female		Male		F-M	Black (B)		White (W)		B-W	Hispanic (H)		White, Non-Hispanic (WNH)		H-WNH
	M	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
WPA-Dimensions															
Realistic (R)	2.97	0.85	3.60	0.73	-0.74	3.12	0.91	3.52	0.77	-0.44	3.48	0.81	3.52	0.76	-0.05
Investigative (I)	3.31	0.67	3.26	0.65	0.07	3.39	0.67	3.24	0.65	0.22	3.40	0.67	3.21	0.63	0.28
Artistic (A)	2.86	0.82	2.76	0.75	0.12	2.93	0.79	2.73	0.76	0.25	2.90	0.80	2.71	0.75	0.24
Social (S)	3.88	0.64	3.54	0.64	0.53	3.84	0.66	3.57	0.64	0.41	3.74	0.63	3.54	0.64	0.32
Enterprising (E)	3.36	0.61	3.37	0.58	-0.02	3.56	0.64	3.32	0.57	0.38	3.48	0.60	3.30	0.56	0.30
Conventional (C)	3.39	0.68	3.18	0.61	0.31	3.55	0.67	3.16	0.60	0.58	3.38	0.65	3.12	0.58	0.40
Average Absolute d					*0.30*					*0.38*					*0.26*
WPA-Facets															
Mechanical (R)	2.51	1.05	3.31	1.00	-0.76	2.87	1.13	3.18	1.05	-0.27	3.16	1.04	3.17	1.05	-0.01
Physical (R)	3.33	0.93	3.81	0.80	-0.52	3.32	0.96	3.78	0.81	-0.48	3.73	0.87	3.78	0.80	-0.06
Critical Thinking (I)	3.79	0.74	3.76	0.72	0.04	3.83	0.73	3.75	0.72	0.11	3.83	0.72	3.73	0.72	0.14
Conduct Research (I)	2.83	0.80	2.77	0.77	0.08	2.96	0.81	2.73	0.76	0.28	2.97	0.81	2.69	0.74	0.35
Artistic Activities (A)	2.47	0.92	2.34	0.85	0.14	2.53	0.90	2.32	0.86	0.23	2.50	0.92	2.30	0.85	0.22
Creativity (A)	3.63	0.89	3.57	0.85	0.07	3.73	0.90	3.55	0.85	0.20	3.68	0.87	3.53	0.85	0.17
Work with Others (S)	3.98	0.70	3.78	0.71	0.29	3.99	0.72	3.79	0.71	0.28	3.94	0.69	3.76	0.70	0.26
Help Others (S)	3.78	0.74	3.31	0.73	0.64	3.69	0.76	3.35	0.74	0.45	3.54	0.74	3.32	0.74	0.30
Prestige(E)	3.90	0.66	3.88	0.66	0.03	3.99	0.69	3.86	0.65	0.19	3.96	0.66	3.85	0.65	0.17
Lead Others (E)	3.56	0.78	3.56	0.73	0.00	3.70	0.79	3.53	0.73	0.22	3.69	0.76	3.51	0.72	0.24
High Profile (E)	2.47	0.91	2.53	0.87	-0.07	2.88	0.93	2.42	0.85	0.49	2.66	0.91	2.39	0.84	0.30
Information Management (C)	2.85	0.92	2.57	0.81	0.30	3.13	0.89	2.52	0.80	0.69	2.84	0.89	2.48	0.77	0.40
Detail Orientation (C)	3.99	0.80	3.85	0.77	0.18	4.03	0.80	3.85	0.77	0.23	4.01	0.77	3.82	0.77	0.25
Clear Procedures (C)	4.05	0.75	3.86	0.75	0.25	4.10	0.76	3.86	0.75	0.32	4.04	0.77	3.83	0.74	0.27
Average Absolute d					*0.24*					*0.32*					*0.22*

Note. *M* = Scale mean for group, *SD* = Scale standard deviation for group; $d = (M_{COMPARISON} - M_{REFERENT})/SD_{REFERENT}$. The referent groups are Males, Whites, and Non-Hispanic Whites; the comparison groups are Females, Blacks, and Hispanics. The WPA yields six dimension and 14 facet scores. The letters in parentheses after the name of the facet scores denotes the higher order dimension.

APPENDIX E
WORK PREFERENCES ASSESSMENT (WPA) ITEM REDUCTION

Matthew T. Allen, Dan J. Putka, and Michael J. Ingerick
(HumRRO)

Background

The WPA was identified as a promising measure for enhancing new Soldier selection. However, limited time would be available to administer the WPA in addition to the TAPAS in an Initial Operational Test and Evaluation (IOT&E). The version of the WPA administered in the Army Class research consisted of 72 items and had a 20-minute time limit. The goal for the IOT&E was to reduce the WPA's administration time to 15 minutes. We considered two objectives when reducing the WPA:

1. Reduce the number of items in the WPA by 15% to 20% (i.e., 10 to 15 items). While a decrease in administration time from 20 to 15 minutes represents a 25% reduction, two factors suggested that deleting 15% to 20% of the items would suffice: (a) the time to complete a measure generally decreases when administered by computer versus paper and pencil (e.g., respondents do not need to spend as much time filling in bubbles), and (b) data on the time taken by Soldiers to complete the WPA during the Army Class concurrent validation data collection indicated that most Soldiers (95%) completed the measure in 15 minutes or less.[9]

2. Equally distribute the number of items eliminated from the WPA across scales. To minimize any loss in the WPA's predictive efficacy from reducing the measure, we elected to drop individual items rather than entire scales. Wherever feasible, our goal was to equally distribute the number of items eliminated across the different scales constituting the WPA.

With these objectives in mind, an analysis approach was formulated to determine which items were the best candidates for elimination.

Approach

Identifying Items for Elimination

An iterative, multi-step approach was taken to identify items for elimination from the WPA. In the first step, one or two items were identified from each scale having more than three items. In the second step, additional items were considered until a total of 10-15 items had been identified for elimination. Four statistical and practical factors were considered when determining which items to drop:

[9] See Ingerick et al. (2008) for details on the Army Class concurrent validation data collection.

1. *Internal consistency*. Items that were unrelated to the other items in the scale, as measured by the item-total correlations, were considered for elimination. Closer scrutiny was given to items that, if eliminated, would raise the scale's item-total correlation.

2. *Item redundancy*. Items that were redundant with other items in the scale, as measured by the bivariate correlations among items in a particular scale, were considered for elimination from the WPA. Only items correlated above .90 were considered for elimination.

3. *Item-level predictive validity estimates*. Items that predicted little to no variance in the criteria of interest, as measured by bivariate and partial (controlling for AFQT) correlations between the items and the criteria, were considered for elimination. Nonsignificant (two-tailed) correlations were scrutinized more closely in considering items to eliminate. Consistent with Chapter 3, five criterion measures that were of value to the Army and comprehensively cover the criterion space were chosen for this evaluation. These criteria were (a) the Army-Wide (AW) Performance Rating Scale (PRS) measuring Effort, (b) the Army Life Questionnaire (ALQ) Army Physical Fitness Test (APFT) score, (c) Attrition (6-month), (d) the ALQ Number of Disciplinary Incidents, and (e) the MOS-Specific Job Knowledge Test (JKT).

4. *Number of items in the scale*. Some WPA facet scales had many more items than others. Scales that contained more items were more closely scrutinized when identifying items for elimination. However, efforts were made to ensure the reductions were equally distributed across as many of the scales as possible. Scales that had three items or fewer were left intact.

Of these factors, the last two received the greatest weight when determining which items to eliminate.

Evaluating the Effects of Reducing the WPA

After a set of 10-15 items were identified for elimination, we evaluated the effects of deleting those items on the predictive validity and internal consistency reliability of the WPA. We accomplished this by comparing the full WPA's incremental validity (over the AFQT) to the incremental validity estimates using the reduced WPA scales. Incremental validity was estimated using OLS (or logistic) regression, where Soldiers' scores on the criterion of interest were regressed on their AFQT scores in the first step and scores on all of the WPA facet scales were entered in the second step. The change in R (ΔR) between the two steps provided an estimate of the WPA's incremental validity. The effects of reducing the WPA on its reliability was evaluated by comparing the estimated coefficient alpha for each scale before and after deleting the targeted set of items. As necessary, further modifications were made to the items selected for elimination based on these results.

Findings

Using the four decision rules described above, 10 items were identified for elimination from the WPA in the first step. Only one item per scale was selected for elimination in this step. Although the 10 items reached the minimum goal for WPA reduction, the remaining items were more closely examined to determine whether any additional items should be dropped. In this second step, three additional items were identified for elimination. In sum, a total of 13 items were identified for deletion. Consistent with our objectives, the number of items deleted ranged from one to two per scale. Table E.1 summarizes the number of items identified for elimination and the reason(s) for their deletion by WPA scale.

Table E.1. Summary of WPA Items Identified for Deletion by Scale

Scale	Number of Items in Original	Number of Items Identified for Deletion	Reason(s) for Deleting Selected Item(s)
Realistic Interests (Dimension)	*13*		
Mechanical (Facet)	5	1	Low item-level predictive validity estimates.
Physical (Facet)	7	1	Low item-level predictive validity estimate for APFT.
Investigative Interests (Dimension)	*12*		
Critical Thinking (Facet)	6	2	Low item-level predictive validity estimates.
Conduct Research (Facet)	6	2	Low item-level predictive validity estimates.
Artistic Interests (Dimension)	*12*		
Artistic Activities (Facet)	8	2	Low item-level predictive validity estimates.
Creativity (Facet)	4	0	
Social Interests (Dimension)	*10*		
Work with Others (Facet)	5	1[a]	Low item-level predictive validity estimates.
Help Others (Facet)	5	1	Low item-level predictive validity estimates.
Enterprising Interests (Dimension)	*13*		
Prestige (Facet)	5	1	Low item-level predictive validity estimates.
Lead Others (Facet)	4	1	Low item-total correlation.
High Profile (Facet)	4	0	
Conventional Interests (Dimension)	*12*		
Information Management (Facet)	6	1	Low item-level predictive validity estimates.
Detail Orientation (Facet) [b]	3	0	
Clear Procedures (Facet) [b]	3	0	
Total	*72*	*13*	

Note. The Realistic dimension scale contains one item that is not in any of the individual facets. This item was retained.
[a] Item was removed originally and then added back in to the reduced version after further analysis.
[b] Scales have overlapping items.

After a set of items for had been identified for elimination, scores on the reduced WPA scales were recomputed and their incremental validity and reliability compared to estimates based on the original, full WPA scales. The results of the incremental validity analyses suggested that the elimination of one item from the Work with Others facet scale had a strong negative impact on the scale's predictive validity. For this reason, this item was added back to the scale. The results of the evaluation analyses after this item was excluded from elimination can be found in Tables E.2 and E.3.

As shown in Table E.2, the elimination of these items had little to no effect on their respective scales' incremental validity. In all cases, incremental validity estimates (ΔR) remained about the same or increased slightly with the reduced versions of the scales. No appreciable explanatory variance was lost due to the elimination of these 12 items.

Similarly, eliminating the 12 items had minimal to negligible effects on the scales' reliability (see Table E.3). Although the coefficient alpha estimates were generally lower for the reduced versions of the scales, the reduction was quite small (less than .05) and remained above .70 in most cases. The lone exceptions to this trend were the findings for the Conduct Research facet (α = .67, .09 drop in value) and the Help Others facet (α = .67, .05 drop in value). However, as Table E.4 demonstrates, these reductions in alpha did not lead to a significant drop in the predictive validity of these facet scales. Therefore, the reduced scales were left as they were.

Table E.2. Incremental Validity Estimates for the Full and Reduced Versions of the WPA over the AFQT

		Full			Reduced			
Criterion/Predictor	n[a]	AFQT Only	AFQT + Predictor	ΔR	AFQT Only	AFQT + Predictor	ΔR	ΔR-Diff
MOS-Specific Job Knowledge Test								
WPA Dimensions [6]	1,050	**.476**	**.501**	**.025**	**.476**	**.507**	**.031**	.006
WPA Facets [14]	1,050	**.476**	**.511**	**.035**	**.476**	**.517**	**.041**	.006
Six-Month Attrition								
WPA Dimensions [6]	2,955	.029	**.124**	**.095**	.029	**.126**	**.097**	.002
WPA Facets [14]	2,953	.029	**.162**	**.133**	.029	**.166**	**.137**	.004
Disciplinary Incidents (Dichotomous)								
WPA Dimensions [6]	1,098	**.095**	.145	.050	**.095**	.141	.046	-.004
WPA Facets [14]	1,098	**.095**	.181	.086	**.093**	.172	.079	-.007
Physical Fitness (APFT)								
WPA Dimensions [6]	1,092	.035	.071	.036	.035	.081	.046	.010
WPA Facets [14]	1,092	.035	**.168**	**.133**	.035	**.190**	**.155**	.022
Effort Ratings Composite (Army-Wide)								
WPA Dimensions [6]	1,128	**.184**	.197	.013	**.184**	.196	.012	-.001
WPA Facets [14]	1,128	**.184**	.213	.029	**.184**	.215	.031	.002

Note. AFQT = Armed Forces Qualification Test. *AFQT Only* = Correlation between the AFQT and the criterion. *AFQT + Predictor* = Multiple correlation (R) between the AFQT and the selected predictor measure with the criterion. ΔR = Increment in R over the AFQT from adding the selected predictor measure to the regression model (AFQT + Predictor – AFQT Only). ΔR-Diff = $\Delta R_{Reduced} - \Delta R_{Full}$. For the dichotomous criteria (6-month attrition and disciplinary incidents), ΔR = Increment in Nagelkerke's R over AFQT. All estimates are uncorrected for statistical artifacts (e.g., range restriction, criterion unreliability). Estimates in bold are statistically significant, $p < .05$ (two-tailed). The numbers in brackets after the title of the predictor measure indicate the number of scale scores that the measure contributed to the regression model.

[a] There were slightly fewer cases included in the reduced WPA facets analyses. This led, in some cases, to slightly different values in the AFQT only statistics.

In sum, a total of 12 items across nine WPA facet scales (17%) were chosen for deletion. Dropping these items did not appear to adversely affect the incremental validity or reliability of the WPA.

Table E.3. Coefficient Alphas for the Full and Reduced Versions of the WPA Scales

Scale	Full	Reduced	Diff
Realistic Interests (Dimension)	.91	.88	-.03
Mechanical (Facet)	.90	.87	-.03
Physical (Facet)	.89	.88	-.01
Investigative Interests (Dimension)	.85	.80	-.04
Critical Thinking (Facet)	.82	.79	-.04
Conduct Research (Facet)	.76	.67	-.09
Artistic Interests (Dimension)	.88	.86	-.01
Artistic Activities (Facet)	.86	.84	-.02
Creativity (Facet)	.82	.82	-.00
Social Interests (Dimension)	.83	.81	-.02
Work with Others (Facet)	.77	.77	-.00
Help Others (Facet)	.71	.67	-.05
Enterprising Interests (Dimension)	.81	.80	-.02
Prestige (Facet)	.68	.64	-.04
Lead Others (Facet)	.73	.73	-.00
High Profile (Facet)	.71	.74	.04
Conventional Interests (Dimension)	.82	.80	-.02
Information Management (Facet)	.82	.77	-.05
Detail Orientation (Facet)	.73	.73	-.00
Clear Procedures (Facet)	.64	.64	-.00

Note. Full = coefficient alphas for scales including all WPA items, Reduced = coefficient alphas for WPA scales excluding the 12 items identified for deletion, Diff = $\alpha_{Reduced} - \alpha_{Full}$.

Table E.4. Standardized Betas for the Full and Reduced Versions of the Conduct Research and Help Others Facet Scales

		Full		Reduced		
Criterion/Scale	n[a]	β	p	β	p	p-diff
MOS-Specific Job Knowledge Test						
Conduct Research (Facet)	1,050	-.016	.677	-.039	.306	-.371
Help Others (Facet)	1,050	-.022	.559	-.022	.550	-.010
Six-Month Attrition						
Conduct Research (Facet)	2,955	-.128	.276	-.150	.186	-.090
Help Others (Facet)	2,953	.338	.003	.354	.001	-.002
Disciplinary Incidents (Dichotomous)						
Conduct Research (Facet)	1,098	-.109	.439	.026	.840	.401
Help Others (Facet)	1,098	-.199	.170	-.216	.116	-.054
Physical Fitness (APFT)						
Conduct Research (Facet)	1,092	-.024	.572	.017	.688	.116
Help Others (Facet)	1,092	.088	.039	.103	.013	-.027
Effort Ratings Composite (Army-Wide)						
Conduct Research (Facet)	1,128	-.024	.573	-.014	.745	.172
Help Others (Facet)	1,128	.043	.304	.032	.428	.124

Note. β = Standardized beta weight from OLS or logistic regression, Full = regression results for scales including all WPA items, Reduced = regression results for WPA scales excluding the items identified for deletion, p-Diff = $p_{Reduced} - p_{Full}$.

[a] There were slightly fewer cases included in the reduced WPA facet analyses.