Technical Report 1311

Development of the Tailored Adaptive Personality Assessment System (TAPAS) to Support Army Selection and Classification Decisions

Fritz Drasgow, Stephen Stark, Oleksandr S. Chernyshenko, Christopher D. Nye, and Charles L. Hulin
Drasgow Consulting Group

Leonard A. White
U.S. Army Research Institute

August 2012

United States Army Research Institute
for the Behavioral and Social Sciences

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

Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

Michelle Sams

MICHELLE SAMS, Ph.D.
Director

Research accomplished under contract
for the Department of the Army

Drasgow Consulting Group

Technical review by

Tonia S. Heffner, U.S. Army Research Institute
J. Douglas Dressel, 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, 6000 6th Street (Bldg. 1464 / Mail Stop 5610), Ft. Belvoir, VA 22060-5610.

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

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

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy) August 2012	2. REPORT TYPE Final	3. DATES COVERED (from. . . to) December 2005 to September 2011

4. TITLE AND SUBTITLE Development of the Tailored Adaptive Personality Assessment System (TAPAS) to Support Army Personnel Selection and Classification Decisions	5a. CONTRACT OR GRANT NUMBER W74V8H-06-C-0006
	5b. PROGRAM ELEMENT NUMBER 622785
6. AUTHOR(S) Fritz Drasgow, Stephen Stark, Oleksandr S. Chernyshenko, Christopher D. Nye, and Charles L. Hulin (Drasgow Consulting Group); Leonard A. White (U.S. Army Research Institute)	5c. PROJECT NUMBER A790
	5d. TASK NUMBER 329
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Drasgow Consulting Group (DCG) 3508 Highcross Rd. Urbana, IL 61802	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 6000 6th Street (Bldg. 1464 / Mail Stop 5586) Fort Belvoir, VA 22060	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Technical Report 1311

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Contracting Officer's Representative and Subject Matter Expert POC: Dr. Leonard White

14. ABSTRACT *(Maximum 200 words)*:

The U.S. Army requires efficient and effective methods for entry-level Army selection and classification decisions. Accordingly, the Tailored Adaptive Personality Assessment System (TAPAS) was developed to assess personality factors related to performance in the Army. TAPAS assesses up to 21 subdimensions of the Big Five personality factors and several additional personality characteristics relevant to military settings. Of particular importance is that TAPAS is designed to be resistant to faking good, so that it can be used for high stakes assessment such as enlistment testing. Each TAPAS item consists of two statements, balanced in social desirability, and a respondent picks the statement that is "more like me." Two item pools were developed and item response theory was used for to administer items as a computerized adaptive test (CAT). Early results from an initial operational test and assessment (IOT&E) indicate little adverse impact on females and minority groups. In addition, mean scores for Army applicants who take TAPAS as part of enlistment screening are very similar to Air Force applicants who are administered TAPAS for research purposes only, which indicates good resistance to faking.

15. SUBJECT TERMS

Personnel, Personality Assessment, Selection and Classification

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited	128	Dorothy Young 703-545-2316

Technical Report 1311

Development of the Tailored Adaptive Personality Assessment System (TAPAS) to Support Army Selection and Classification Decisions

Fritz Drasgow, Stephen Stark, Oleksandr S. Chernyshenko, Christopher D. Nye, and Charles L. Hulin
Drasgow Consulting Group

Leonard A. White
U.S. Army Research Institute

Personnel Assessment Research Unit
Tonia S. Heffner, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
6000 6th Street, Bldg. 1464
Fort Belvoir, Virginia 22060

August 2012

Army Project Number	**Personnel, Performance**
622785A790	**and Training Technology**

Approved for public release; distribution is unlimited

ACKNOWLEDGEMENT

We thank the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) for funding aspects of the research described in this report via SBIR Grant A04-29. The authors are especially thankful to Drs. Tonia Heffner and Michael Rumsey.

We also acknowledge the support provided by LTG Benjamin Freakley, Commanding General, U.S. Army Accessions Command.

We also wish to especially thank the thousands of Soldiers who participated in this research; without their cooperation, the work could not have been accomplished.

We are indebted to staff at the Human Resources Research Organization for collecting data in the Expanded Enlistment Eligibility Metrics (EEEM) study.

For implementation of TAPAS at the Military Enlistment Processing Stations, thanks are due to staff at the Defense Manpower Data Center, including Dr. Dan Segall and Ms. Kathy Moreno, the Military Enlistment Processing Command, including Mr. Don Hill and Mr. Brad Tiegs, and the U.S. Air Force, including Ms. Lisa Mills and Dr. Tom Carretta.

DEVELOPMENT OF THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM (TAPAS) TO SUPPORT ARMY SELECTION AND CLASSIFICATION DECISIONS

EXECUTIVE SUMMARY

Research Requirement:

The U.S. Army requires efficient and effective methods for selecting new Soldiers. To this end, the Army utilizes the Armed Services Vocational Aptitude Battery (ASVAB) to identify applicants who satisfy enlistment requirements. The ASVAB has proven to be very useful and will continue to play an important role in selecting new Soldiers. However, additional personal attributes, in particular non-cognitive characteristics, are important for entry-level Soldier performance and retention (Drasgow, Embretson, Kyllonen, & Schmitt, 2006).

The Tailored Adaptive Personality Assessment System (TAPAS) was developed to assess some of these non-cognitive characteristics. Based on recent research on item response theory (IRT), computerized adaptive testing (CAT), and temperament/personality measurement, TAPAS provides a means for validly and efficiently assessing personality in a way that is fake resistant. The research had three primary components. First, Stark's (2002; Stark, Chernyshenko, & Drasgow, 2005) multidimensional pairwise preference (MDPP) model provides the psychometric theory for an item format that is resistant to socially desirable responding or, more generally, faking. Second, using the MDPP model, a new approach to computer adaptive testing was developed. It incorporates new methods for test construction and scoring, using the MDPP format. Third, the TAPAS trait taxonomy was developed to provide a comprehensive but non-redundant description of the lower order personality dimensions (which are often called facets) that underlie the Big Five personality model. The TAPAS trait taxonomy includes 22 facets of personality, and has since been expanded to 28 dimensions by the addition of facets of particular interest in military settings.

Procedure:

An IRT approach to constructing and scoring MDPP items was developed and evaluated. With this approach, a CAT algorithm was developed. The TAPAS trait taxonomy was devised and pools of personality statements to measure the TAPAS facets were written. These items were administered to large samples of new Soldiers in reception battalions and their IRT item parameters were estimated. To better understand the predictive value of the TAPAS facets, a data-base consisting of over 1,600 correlations between personality dimensions and eight dimensions of performance was created. In addition, a laboratory study and a field study were conducted. Simulation studies were conducted to evaluate the CAT algorithm for TAPAS. TAPAS was successfully implemented at six Military Enlistment Processing Stations (MEPS) in June 2009 and expanded to all MEPS in September of that year. A second item pool, to be used exclusively for enlistment screening, was also developed.

Findings:

Estimation of IRT item parameters yielded many items with satisfactory item discrimination. The simulation studies found that using CAT can cut the number of items needed to reach a given level of precision by approximately 50% and therefore greatly reduce testing time. Early findings from this initial operational testing and evaluation (IOT&E) provide evidence for the construct validity of TAPAS facets. Importantly, only small differences in mean scores have been found in comparisons across race and gender groups. In addition, very little difference in mean scores has been found for Army applicants, who take TAPAS as part of enlistment screening, and Air Force applicants, who take TAPAS for research purposes only. This is powerful evidence for TAPAS's resistance to faking.

Utilization and Dissemination of Findings:

These findings have been presented to senior leaders in the Army. Specifically, TAPAS provides a personality assessment system with the flexibility and robustness needed for assessment in high stakes settings in which examinees are aware that their scores will be used to make important decisions about them. The development of TAPAS has required innovations in psychometric theory, computer adaptive testing technology, and personality theory. Together, these developments have led to a flexible and fake-resistant approach to personality assessment that holds great promise for improved enlistment screening.

DEVELOPMENT OF THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM (TAPAS) TO SUPPORT ARMY SELECTION AND CLASSIFICATION DECISIONS

CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

DEVELOPMENT OF THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM (TAPAS) TO SUPPORT ARMY SELECTION AND CLASSIFICATION DECISIONS

CHAPTER 1: INTRODUCTION TO TAPAS AND THE MULTI-UNIDIMENSIONAL PAIRWISE PREFERENCE IRT MODEL

This chapter provides the rationale for the new Tailored Adaptive Personality Assessment System (TAPAS) and describes a sequence of psychometric studies that demonstrate why multidimensional pairwise preference items should be viewed as a promising alternative for dealing with the faking problem. Specifically, we briefly summarize our item response theory (IRT) approach to constructing and scoring multidimensional pairwise preference items and present the initial simulation results obtained by Stark (2002) showing that accurate normative scores can be obtained. We then discuss additional studies conducted to show that our proposed approach can produce accurate normative scores in as many as 5-dimensions with minimal inter-dimensional linking requirements. We conclude the chapter by suggesting that the introduction of a formal psychometric model paves a way for computerized adaptive personality testing, which is particularly attractive in operational testing contexts.

Why TAPAS?

Interest in temperament/personality as a predictor of work performance has increased considerably over the last twenty years. This increase has been caused, in part, by legal and societal concerns about adverse impact associated with the use of intelligence test scores for selection and promotion. Interest has also been stimulated by empirical evidence showing that personality constructs predict performance across a diverse array of civilian and military occupations (e.g., Barrick & Mount, 1991; Campbell & Knapp, 2001) and provide incremental validity beyond general cognitive ability (Schmidt & Hunter, 1998).

The use of personality variables for selection and classification inevitably calls attention to the quality of their measurement. Historically, personality researchers have been more concerned with developing trait taxonomies and examining predictive validity than with exploring and developing new methods for scale construction and scoring. Even today, the majority of research focuses on whether broad personality dimensions, such as the Big Five personality factors, are better for predicting job-related outcomes than narrower, lower-order facets (e.g., Paunonen & Jackson, 2000), but relatively little research examines the potential benefits of using modern psychometric (i.e., item response theory [IRT]) based methods to construct, score, and administer personality items. Thus, for the most part, personality batteries still consist of many scales having ten or so single stimulus (statement) items that are administered with a dichotomous (Agree/Disagree) or polytomous (i.e., Likert) response format.

Although traditional personality measures are undoubtedly useful in research and counseling settings, where respondents are inclined to answer honestly, they may be less appropriate for making important personnel decisions for several reasons. First, in high stakes testing situations, research shows that single statement personality items can be easily faked: Test takers can discern the socially desirable answers, which are scored as "correct," and, thus, increase or decrease their scores to suit their personal needs (White & Young, 1998; White, Young, Hunter, & Rumsey, 2008). This intentional distortion can severely undermine the utility of measures for personnel selection. Second, because classical test theory methods have been used to evaluate and choose items during the development of almost all currently available personality scales, only those items with moderately positive or moderately negative standings on the underlying trait continuum were retained; extreme and neutral items were discarded (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Stark, Chernyshenko, Drasgow, & Williams, 2006). This degrades the validity of the scale for identifying the rank-order of the high and low scoring individuals who are often of primary interest in selection contexts. Third, traditional personality measures have usually only one or, at best, two test forms, which decreases test security through repeated item exposure (Guo, Tay, & Drasgow, 2009).

These concerns are not mere academic quibbles. In the early 1990s, Navy researchers sought to implement a single statement personality measure called the Armed Services Applicant Profile (ASAP) and Army researchers had similar intentions for an instrument called the Assessment of Background and Life Experiences (ABLE). An impressive set of studies showed that these measures predicted important behaviors (e.g., Trent & Quenette, 1992; White, Nord, Mael, & Young, 1993; White, Young, & Rumsey, 2001). Nonetheless, the Department of Defense Advisory Committee on Military Personnel Testing (DAC-MPT) did not recommend implementation because of concerns that single statement items were easily compromised by faking good (White et al., 1993).

Keeping these concerns in mind, as well as the advantages that modern psychometric methods and computing technology can offer, there was a unique opportunity to develop a new generation of personality measures that 1) are more *fake-resistant*, 2) use *computer adaptive* technology to measure well across a broad range of trait continua, and 3) are *easily customized* to meet the needs of many different organizations. The resulting measure, called the Tailored Adaptive Personality Assessment System or TAPAS, represents the confluence of research efforts in the areas of item response theory, computerized adaptive testing, and personality/personality measurement. TAPAS utilizes a multidimensional pairwise preference (MDPP) item format that is intended to reduce faking. MDPP items are constructed from pools of precalibrated personality statements that measure dimensions relevant to performance in the military. Because TAPAS test construction and scoring is based on a formal IRT model (Stark, 2002; Stark et al. 2005), both nonadaptive (static) and adaptive tests can be built to measure many dimensions simultaneously with a good balance between measurement precision and testing time.

Psychometric Theory for Fake-Resistant Personality Assessment

Many organizations are reluctant to rely on personality scores for making important personnel decisions because of concerns about faking, commonly defined as intentional response distortion. Research using traditional paper and pencil personality measures indicates that faking can affect the overall factor structure of test batteries, correlations among subscales, criterion-related validity, utility of top-down selection systems, and test scores based on both classical and traditional unidimensional IRT methods (for a detailed review, see Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001**)**. Furthermore, efforts to “correct” for faking post hoc using social desirability or impression management scores have been largely unsuccessful. Corrections for faking have had little salutary effect on validity or utility (Ellingson, Sackett, & Hough, 1999), perhaps because they partial out variance related to performance (Ellingson et al., 1999; Kriedt & Dawson, 1961), or because scales designed to detect socially desirable responding are also susceptible to response distortion (Stark et al., 2001; Zickar & Drasgow, 1996). Consequently, there has been an attempt to reduce faking through strategic item construction. Rather than presenting items consisting of individual statements that describe how one typically thinks, feels, or acts, and asking respondents to indicate their levels of agreement on a scale of, say, 1 (Strongly Disagree) to 5 (Strongly Agree), items can be presented in the form of multidimensional forced choice blocks involving two to five statements that have been matched on social desirability. By asking respondents to indicate their relative preferences via ranks or to select the statement that is "most like me" and the statement that is "least like me," one can readily obtain information that permits intra-individual score comparisons. However, because personnel selection applications require normative scores, new methods were needed to circumvent the ipsativity problems (Hicks, 1970) traditionally associated with multidimensional forced choice measures.

Fortunately great strides have been made in the area of multidimensional forced choice testing in recent years. The classical test theory methods that were developed for the Assessment of Individual Motivation (AIM; White & Young, 1998) have proven effective for producing normative data needed for selection purposes (e.g., Jackson, Wrobleski, & Ashton, 2000; McCloy, Heggestad, & Reeve, 2005). By varying the number of statements representing each personality dimension and by differentially weighting response alternatives, researchers have successfully created scales that are only “partially ipsative” (Hicks, 1970), meaning that there is enough variation in the total scores, formed by summing points over all dimensions, to permit meaningful inter-individual comparisons (e.g., White & Young, 1998). In addition, great progress has been made in the IRT realm. Stark et al. (2005) proposed a method for constructing and scoring multidimensional pairwise preference tests that was effective in recovering trait scores in two-dimensional computer simulations, and since then simulation studies have shown good to excellent recovery of trait scores with tests involving up to 25 dimensions (Stark, Chernyshenko, & Drasgow, 2011). Moreover, this methodology has been shown to produce valid trait scores in laboratory experiments (Chernyshenko et al., 2009), Army field studies involving job incumbents (Knapp & Heffner, 2010), and as we review later in this report, trait scores showing stable means, intercorrelations, and validities in operational testing settings (Knapp, Heffner, & White, 2011) Moreover, the multidimensional pairwise preference model proposed by Stark et al. was extended recently by de la Torre and colleagues to more complex item formats, and their independent studies not only validated the psychometric tenets of the model but provided supporting evidence for trait score recovery (De la Torre, Ponsoda, Leenen, & Hontangas, 2011).

An IRT Approach to Constructing and Scoring Multidimensional Pairwise Preference Items

The model proposed by Stark (2002) assumes that when a respondent is presented with a pair of stimuli (e.g., two personality statements), denoted *s* and *t*, and is asked to indicate a preference, he or she evaluates each statement separately and makes independent decisions about statement endorsement. If a respondent's endorsement propensity is equal for both statements, he or she must reevaluate the statements independently until a preference is reached, as indicated by endorsing one statement and not endorsing the other. Thus, the probability of preferring statement *s* to statement *t* in item *i*, given trait scores $(\theta_{d_s}, \theta_{d_t})$ on the dimensions, d_s and d_t represented by those statements, can be written as

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

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 pairwise preference items (i.e., pairings), where *i* = 1 to *n*;

$d =$ index for dimensions, where *d* = 1, …, *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;

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

and

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

With the assumption that the two statements in each pairwise preference item 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 obtain the desired form of the equation for MDPP response probabilities:

$$(2) \qquad P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P_s(1 \mid \theta_{d_s}) P_t(0 \mid \theta_{d_t})}{P_s(1 \mid \theta_{d_s}) P_t(0 \mid \theta_{d_t}) + P_s(0 \mid \theta_{d_s}) P_t(1 \mid \theta_{d_t})},$$

where

$P_s(1 \mid \theta_{d_s}) =$ probability of endorsing statement *s* given trait score θ_{d_s};

$P_s(0 \mid \theta_{d_s}) =$ probability of not endorsing statement *s* given trait score θ_{d_s};

$P_t(1 \mid \theta_{d_t}) =$ probability of endorsing statement *t* given trait score θ_{d_t};

and

$P_t(0 \mid \theta_{d_t}) =$ probability of not endorsing statement *t* given trait score θ_{d_t}.

The probability of endorsing a stimulus in a pairwise preference item depends on θ_{d_s} and θ_{d_t} and also depends fundamentally on the model chosen to characterize the process of stimulus responding. In principle, any IRT model for unidimensional single stimulus responses could be chosen for computing the $P_s(1|\theta_{d_s})$ and $P_t(1|\theta_{d_t})$ terms in Equation 2. However, research suggests that *ideal point* models should be used (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Drasgow, Chernyshenko, & Stark, 2010; Stark et al., 2006) because they provide a better representation of the process by which people respond to personality statements. Whereas dominance models tend to highlight statements in a pool having high item-total correlations and linear factor loadings, ideal point models not only identify those but also discriminating statements that reflect positions of neutrality or moderation. Consequently the pool of stimuli available for MDPP test construction is expanded when using an ideal point model for statement calibration. For TAPAS, we utilized the Generalized Graded Unfolding Model (GGUM; Roberts, Donoghue, & Laughlin, 2000), because it is one of the most flexible ideal point models developed to date and it has been shown to fit data for individual personality statements well in our previous investigations (Chernyshenko et al., 2009; Stark et al., 2006).

Once MDPP tests have been constructed and administered, the scoring of response patterns can be accomplished, for example, by Bayes modal estimation. For a vector of latent trait scores, $\tilde{\boldsymbol{\theta}} = (\theta_{d'=1}, \theta_{d'=2}, ..., \theta_{d'=D})$, this involves maximizing:

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

where $\tilde{\mathbf{u}}$ represents an examinee's response pattern, u_i is the dichotomous response to item *i*, $P_{(s>t)_i}$ is the probability of preferring statement *s* to statement *t* in item *i*, and $f(\tilde{\boldsymbol{\theta}})$ is a *D*-dimensional prior density function, which, for simplicity, is assumed to be the product of independent normals,

$$\prod_{d'=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\theta_{d'}^2}{2\sigma^2} \right).$$

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

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

leaving the following set of equations to be solved numerically:

$$\frac{\partial \ln L}{\partial \tilde{\boldsymbol{\theta}}} = \begin{bmatrix} \dfrac{\partial \ln L}{\partial \theta_{\mathrm{d'=1}}} \\ \dfrac{\partial \ln L}{\partial \theta_{\mathrm{d'=2}}} \\ \ldots \\ \dfrac{\partial \ln L}{\partial \theta_{\mathrm{d'=D}}} \end{bmatrix} = 0.$$

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 that DFPMIN does not require an analytical solution for the second derivatives, which would require extensive and complicated calculations.

The Initial Study Examining Trait Score Recovery with the Proposed MDPP Test Construction and Scoring Procedures

We first conducted simulation studies where the recovery of known trait scores was investigated for one-dimensional (1-D) and two-dimensional (2-D) pairwise preference tests, as described by Stark (2002). Using dichotomized responses to personality statements that had been administered to Army recruits under instructions to answer honestly, we separately estimated GGUM stimulus parameters for the statements from each dimension. Next, 1-D and 2-D pairwise preference tests were assembled. Items for these tests were formed by selecting pairs of statements with similar social desirability ratings, which had also been collected from Army recruits under instructions to fake good,. Because little was known about how dimensionality would affect test length requirements for trait estimation, a rule of thumb was developed for constructing forms: Tests would be created using 10, 20, or 40 "*items per dimension.*" Thus, 1-D tests based on *10 items per dimension* would comprise 10 pairwise preference items. 2-D tests involving *10 items per dimension* would comprise 20 pairwise preference items, and so on. Following this rule, 1-D pairwise preference tests of 10, 20, and 40 items were constructed by pairing statements that were as similar as possible in desirability, but differed by at least one in their GGUM IRT item extremity parameters. (This was necessary to produce items that were informative in a psychometric sense.)

To create 2-D pairwise tests, statements were matched on both social desirability and their item extremity parameter. In addition, it was necessary to include a proportion of unidimensional pairings to fix the scale (i.e., to allow normative scores to be estimated, rather than ipsative scores). But, because this was the first study to explore multidimensional pairwise preference test construction and scoring, the proportion of unidimensional pairings was included as a design factor. By combining three levels of test length (20, 40, and 80 pairwise preference items) and three proportions of unidimensional pairings (10%, 20%, and 40%), nine 2-D tests were produced.

The resulting 1-D and 2-D tests were "administered" to large numbers of simulated examinees whose latent trait scores were known; responses to the pairwise preference items were generated based on the multi-unidimensional pairwise preference model, shown in Equation 2. Test scores for the simulated respondents were estimated using a multidimensional Bayes modal algorithm that was developed in Excel Visual Basic for Applications (VBA), and the test scores were compared to the known values using statistical indices of trait score recovery.

Results

The 1-D simulation was conducted to establish a baseline for scoring accuracy. Here the correlations between estimated and generating latent trait scores ranged from .90 (10 pairwise preference items) to .97 (40 items). In the 2-D simulation, which had direct implications for application of this method to the problem of faking, the correlations ranged from .77 in the most unfavorable condition (a 20-item test with 10% unidimensional pairings) to .96 in the most favorable (80-item test with 40% unidimensional pairings). The average correlation between estimated and known trait scores was about .9 for the 40-item tests, regardless of the proportion of unidimensional pairings, and the correlations would likely have been higher if more items could have been constructed to provide information at the extremes of the trait continua. Importantly, these results suggested that adequate score recovery could be achieved using tests of 20 or fewer items per dimension and 20% or fewer unidimensional pairings. These promising results provided a basis for exploring tests of higher dimensionality.

Simulation Studies Examining Linking Designs and Trait Score Recovery with 3-D and 5-D Tests

The one- and two-dimensional simulations described by Stark et al. (2005) demonstrated the viability of the MDPP approach to test construction and scoring. However, key questions remained as to the numbers and types of pairings that would be needed to maintain estimation accuracy with tests involving higher dimensionality. In field applications, we envisioned that tests of as many as 15 dimensions might be desired for predicting an array of organizational criteria and thwarting attempts to fake, but exploring so many dimensions at the outset would have been impractical because too little was known about test construction practices and scoring efficacy. Consequently, our Phase I psychometric research addressed the more tractable problem of constructing and scoring MDPP tests of up to five dimensions and generalizing Stark's (2002) Visual Basic for Applications (VBA) scoring program (VBA was used because it is highly flexible and tightly integrated with other Microsoft software products). The findings of these simulations informed the subsequent development of the nonadaptive MDPP test, known as TAPAS-95s (Stark, Chernyshenko, & Drasgow, 2010), which was administered in the Expanded Enlistment Eligibility Metrics (EEEM) research (Knapp & Heffner, 2010), as well as a computerized adaptive item selection algorithm and the first TAPAS prototype written in Visual Basic .NET (VB.NET).

In the next two sections we present the results of simulation studies that examined score recovery with 1-D, 3-D, and 5-D tests involving different "linking" designs. By "linking design," we mean a test composition strategy that identifies the scoring metric so as to yield normative information for selection applications. We examined two approaches representing extremes in the range of possibilities: 1) *Complete linking*, in which all possible combinations of dimensions appear in a test, and 2) *Minimal linking*, a "bare-bones" approach in which the fewest possible combinations of dimensions are utilized. Sampling broadly from the domain of possible item types has clear advantages for maintaining an examinee's interest, utilizing a pool of statements, and controlling item exposure across examinees, but if all possible combinations are needed to effectively recover trait scores, then test length would become a prohibitive factor when many dimensions are assessed. In addition, the feasibility of minimal linking has implications for CAT. The quicker one can identify the metric (i.e., fix the scale), the sooner response data can be scored and items can be selected based on IRT item information functions.

Complete Linking Simulations

Simulations were conducted using GGUM parameters for personality statements measuring three lower-order dimensions (facets) of the Big Five factor Conscientiousness (Order, Traditionalism, and Responsibility) and two facets of Extraversion (Energy and Dominance). To assess the potential effects of differences in the quality of parameters across facets, we began with 1-D simulations involving 10- and 20- item tests for each personality facet. (Note that 40-item tests were not examined because the previous studies suggested that 20 items per dimension would provide adequate trait score recovery.) Items were constructed in the same manner as described above. However, rather than comparing estimated and generating scores for large numbers of simulees at designated points on a grid representing levels of theta (trait scores), a sampling approach was used instead to manage run times. Specifically, for each 1-D test, 1,000 trait scores were sampled randomly from an independent standard normal distribution; responses were generated based on the multi-unidimensional pairwise preference model; and response patterns were scored using the VB.NET program. The quality of normative score recovery in each condition was then assessed using the correlation between the estimated and known trait scores.

A similar approach was used to develop tests involving three and five personality facets. For the 3-D simulations, we created items by pairing statements representing Order, Traditionalism, and Energy. For the 5-D simulations, we also included Responsibility and Dominance.   Combining levels of three independent variables: 1) Items per dimension (10, 20), 2) Proportion of unidimensional pairings (10%, 20%), and 3) Test Dimensionality (3-D, and 5-D), eight tests in total were created. Table 1.1 shows their composition. Entries in the column labeled “Item Type” indicate the combinations of dimension represented by the pairwise preference items. In Table 1, 1 = Order, 2 = Traditionalism, 3 = Energy, 4 = Responsibility, and 5 = Dominance. For example, Item Type 1-1 represents a pairwise preference item involving two statements measuring Order; Item Type 2-3 represents an item involving statements measuring Traditionalism and Energy, and so on.

The values in columns two through five indicate the frequency with which each item type appeared in a particular test. For example, Item Type 1-2 indicates that a statement representing dimension 1 was paired with a statement representing dimension 2 to form a pairwise preference item. Item Type 1-2 appeared 9 times in the 3-D, 30-item test, involving 10% unidimensional pairings, and 16 times in the 3-D, 60-item test involving 20% unidimensional pairings. Note that the frequencies in the respective columns 2 through 5 add up to either 30 or 60 as shown in the table headers. The 3-D tests consisted of either 30 or 60 pairwise preference items, and the 5-D tests consisted of either 50 or 100 pairwise preference items. The percentages (10%, 20%) indicate the proportions of unidimensional items that were included in the respective tests to fix the scale. Importantly, it can be seen that *all possible unique item types* were included in the tests in these complete linking conditions.

Table 1.1. Item Types Appearing in the 3-D and 5-D Tests Using a Complete Linking Design

Item Type	3-D Tests			
	30 Items		60 Items	
	10%	20%	10%	20%
1-1	1	2	2	4
2-2	1	2	2	4
3-3	1	2	2	4
1-2	9	8	18	16
1-3	9	8	18	16
2-3	9	8	18	16

Item Type	5-D Tests			
	50 Items		100 Items	
	10%	20%	10%	20%
1-1	1	2	2	4
2-2	1	2	2	4
3-3	1	2	2	4
4-4	1	2	2	4
5-5	1	2	2	4
1-2	5	4	9	8
1-3	4	4	9	8
1-4	5	4	9	8
1-5	4	4	9	8
2-3	5	4	9	8
2-4	4	4	9	8
2-5	4	4	9	8
3-4	5	4	9	8
3-5	4	4	9	8
4-5	5	4	9	8

In an effort to adequately cover the many possible combinations of trait scores in the 3-D and 5-D studies, 3,000 (3-D) and 5,000 (5-D) trait scores were sampled for the facets in each study from independent standard normal distributions, and the correspondence of estimated and known trait scores was examined using Pearson correlations. Then, to obtain a single index of recovery for each experimental condition, the correlations were averaged across dimensions.

Table 1.2 presents the correlations between known and estimated trait scores across each personality facet and test type. For example, the .95 in the first row of the last column represents the correlation between the estimated and known trait scores for Order, as measured by the 100 item 5-D test with 20% unidimensional pairings. As in the 2-D studies, there was little, if any, effect for the percent of unidimensional pairings, which suggests that 10% unidimensional pairings is all that is required with a complete linking design. This was very desirable in terms of making the proposed TAPAS measures fake resistant. In addition, the correlations between estimated and known trait scores were high even for the short tests in each condition (.88), and they improved with test length. In terms of the number of pairings involving each dimension, the MDPP scoring algorithm seemed to perform better for the 3-D and 5-D tests than for 1-D tests of comparable length, suggesting that the recovery of normative scores might actually improve with higher dimensionality, provided that enough combinations of dimensions are represented by the multidimensional pairings.

Table 1.2. Correlations between Estimated and Known Trait Scores for 1-D, 3-D, and 5-D Tests in the Complete Linking Simulations

Personality Facet	1-D Tests		3-D Tests				5-D Tests			
	10 Items	20 Items	30 Items		60 Items		50 Items		100 Items	
			10%	20%	10%	20%	10%	20%	10%	20%
Order	.88	.93	.91	.91	.95	.95	.91	.91	.95	.95
Traditionalism	.92	.95	.89	.90	.93	.94	.89	.89	.94	.94
Energy	.85	.91	.86	.85	.92	.92	.85	.85	.92	.92
Responsibility	.85	.93	*	*	*	*	.86	.85	.92	.92
Dominance	.90	.93	*	*	*	*	.90	.90	.95	.95
Average	.88	.93	.89	.89	.93	.94	.88	.88	.94	.94

Note. * = data not simulated for this facet; 10% = 10 percent of the pairwise preference items were unidimensional; 20% = 20 percent of items unidimensional.

Examining Minimal Cross-dimensional Linking with 5-D Tests

Because the long term goal for TAPAS was computerized adaptive personality testing, we wanted to examine the relative performance of two approaches for fixing the scale for tests of high dimensionality. In particular, we wanted to determine whether estimation accuracy would diminish markedly if a minimal linking design was used in place of complete linking. Although a variety of minimal linking designs are possible, we chose one that can be referred to as "circular" linking, because we pair dimensions according to their proximity when arranged in a circle. For example, with a 5-D test, this method involves the following item types: 1-2, 2-3, 3-4, 4-5, 5-1, and a designated proportion of unidimensional item types (1-1, 2-2, 3-3, 4-4, and 5-5); the unidimensional item types can be chosen randomly or on substantive grounds, such as their susceptibility to faking.

To compare score recovery with the complete and circular linking designs, we conducted 5-D simulations using the same data generation process, scoring, and method of analysis described above. 5-D tests of 50- or 100- pairwise preference items were created using the appropriate multidimensional item types and either 10% or 20% unidimensional item types were included to identify the metric. The average correlations across dimensions between the estimated and known trait scores are presented in Table 1.3.

Table 1.3. Average Correlations between Estimated and Known Trait Scores for Circular and Complete Linking Designs

Personality Facet	5-D Tests with Circular Linking				5-D Tests with Complete Linking			
	50 Items		100 Items		50 Items		100 Items	
	10%	20%	10%	20%	10%	20%	10%	20%
Order	.91	.92	.96	.95	.91	.91	.95	.95
Traditionalism	.89	.89	.93	.94	.89	.89	.94	.94
Energy	.87	.86	.92	.92	.85	.85	.92	.92
Responsibility	.86	.85	.93	.92	.86	.85	.92	.92
Dominance	.89	.89	.95	.95	.90	.90	.95	.95
Average	.88	.88	.94	.94	.88	.88	.94	.94

Note. The percentages refer to the proportions of items out of 50 or 100 that were unidimensional.

As shown in Table 1.3, the average correlations in the circular and complete linking conditions were identical and the results were remarkably similar at the level of the individual personality facets. This is important because it indicates that circular linking allows for accurate trait score recovery with the MDPP procedure and could be used, if necessary to assess several dimensions with a limited number of item types. Of course, it would be better for many reasons to sample item types broadly from the range of possibilities. However, in situations where faking is a concern, it is clearly *not necessary* to include all possible combinations to recover normative information.

Summary and Conclusions

Together, the simulation results demonstrate that the proposed IRT approach to test construction and scoring can adequately recover normative trait scores in tests comprised of MDPP items. Investigations of test length and linking designs indicated that the method should perform well with tests of 30-100 items, measuring 3 to 5 dimensions, and with only 10% of the item pairs being unidimensional. These results clearly showed that the development of an operational measure based on MDPP scoring would be technically feasible.

The use of formal IRT models for constructing and scoring MDPP items allows for a rigorous evaluation of the psychometric quality of individual statements as well as the items created by pairs of statements. Of critical importance for this evaluation is the *information function* that quantifies how much psychometric “information” is provided by an item at a particular combination of trait levels. This property is fundamental to the concept of adaptive testing, where items are selected to provide near maximal information for each examinee during a test. Specifically, each examinee *receives an individually tailored sequence of items* that can be created on the fly from a large pool of statements that vary in extremity. The computer can search through the pools of statements to find the pairing that optimally assesses a particular individual’s trait levels and this pair of statements will constitute the next item that is presented. Constraints can be imposed on how the statements representing various dimensions are combined and stopping rules can be adopted to terminate testing as soon as the measurement errors in an examinee’s trait scores fall below acceptable thresholds or, alternatively, until a preset number of items has been administered. With CAT, fewer items are needed to achieve the same level of precision as provided by even a well constructed nonadaptive test, so testing time can be reduced or, conversely, more dimensions can be assessed, depending on the need. Computerized testing can also increase test security by imposing “exposure controls” that limit how often individual statements or items are presented to different examinees. The description of the CAT algorithm used in TAPAS as well as results of the simulation studies investigating score recovery and efficiency gains are described in Chapter 5 of this report. The actual implementation of adaptive testing and empirical results is described in Chapter 6.

CHAPTER 2: DEVELOPMENT OF THE TAPAS TRAIT TAXONOMY

This chapter is concerned with the trait domains assessed by the Tailored Adaptive Personality Assessment System (TAPAS). First, we discuss research demonstrating that narrow personality traits are often better predictors of job-related outcomes than broad, higher-order traits. Second, we describe our research efforts to empirically derive a taxonomy of non-redundant narrow personality traits (facets) to form the basis of TAPAS. The resulting 22 facets can be organized hierarchically into five broad personality dimensions: Conscientiousness, Openness to Experience, Extraversion, Agreeableness, and Emotional Stability.

Background

The Big Five theory of personality (Digman, 1990; Goldberg, 1990; Norman, 1963) brought order to the chaotic research literature examining the relation of personality to job performance. Before the Big Five, there was little agreement concerning the basic dimensions of normal personality. The result was a proliferation of instruments that conceptualized personality dimensions in unique ways. One consequence of this was that early studies attempting to combine validities of various personality instruments from different studies in informal meta-analytic ways found near zero correlations with important work outcomes (e.g., Guion & Gottier, 1965). These grim empirical findings were accompanied by the theoretical arguments of Mischel (1968, 1969, 1973) and his colleagues who contended that the behavior of individuals was not sufficiently consistent across time and situations to allow valid predictions by means of personality measures. Fortunately, in the last 20 years, personality researchers have reached a consensus that the Big Five personality factors, Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience, are sufficient to adequately describe normal personality (see Costa & McCrae, 1988; Goldberg, 1990, 1993; Hogan, 1991; Saucier, 1992). This parsimonious factor representation allowed the results of studies on ostensibly different traits to be pooled and, thus, provided a framework for integrating findings from diverse research programs. Meta-analyses by Barrick and Mount (1991), Tett, Jackson, and Rothstein (1991), and others have shown that the Big Five factors are useful in predicting successful performance in many occupational groups, with validities ranging from .2 to .3. And Conscientiousness is especially important because it has been shown to be a consistently valid predictor across all the criterion groups that have been examined. The within person variance in behaviors emphasized by Mischel, thus, represents a limitation on the validity of any set of static predictions, but the consistency of behaviors allows valid predictions of performance across situations.

Although the Big Five theory, and supporting empirical research, constitutes a major contribution to personality theory, it may oversimplify some relationships that are important for personality assessment and the prediction of job performance. Current research suggests that lower-order personality traits are, perhaps, more useful than broad (global) factors from both a theoretical and applied point of view. Of critical importance is that measures of narrow, lower-order traits have been found to have higher predictive validities than measures of broad factors in many recent studies. For example, Paunonen (1998) correlated several Big Five factor and narrow trait measures with various behavioral criteria and concluded that "aggregating personality traits into their underlying personality factors could result in decreased predictive accuracy due to the loss of trait-specific but criterion-valid variance" (p. 538); other researchers have reached similar conclusions (e.g., Ashton, 1998; Mershon & Gorsuch, 1988). Moreover, as noted by Saucier and Ostendorf (1999), there are advantages to using narrow trait measures besides gains in predictive validity. Namely, the use of lower-order traits supports theory development because it clarifies the conceptualization of broader factors (Briggs, 1989). Narrow traits also offer higher fidelity in personality description, and, thus, enhance the diagnostic value of assessment. This is especially beneficial for respondents who fall in the middle of the distribution on a measure of a broad factor, because such scores can be obtained in many different ways. Unlike extreme scores on a broad factor, which suggest that an individual is generally high or low on all subcomponents, middle scores can be attained by being average on all lower-order components or high on some and low on others. This leads to ambiguity in score interpretation and possibly diminishes predictive efficacy.

We note here that this result of validities favoring the use of narrow measures appears to contradict the Bandwidth-Fidelity Paradox (Cronbach & Gleser, 1959): Given a fixed number of items, one can measure a narrow construct with high fidelity or a broad construct with low fidelity. The paradox results because lower fidelity, broader measures typically generate higher validities when used to predict broad performance or behavioral criteria. The paradox discussed by Cronbach and Gleser, and amplified by Humphreys (1970, 1981, 1985, 1986) and Roznowski and Hanisch (1990), was noted in the context of cognitive ability testing, where measures exhibit strong positive intercorrelations (i.e., a positive manifold) due to the presence of a higher order general factor. In contrast the correlations among personality facets, across as well as within the Big Five factors, and the correlations among the Big Five factors themselves tend to be much lower. Thus, narrow facet scores may provide the differential validity for organizational criteria that was once imagined in the cognitive ability domain, and the composites formed by combining narrow trait scores may provide substantial incremental validities for broad organizational outcomes.

Our Empirical Approach to Deriving a Lower-Order Trait Taxonomy for TAPAS

Because of these aforementioned benefits and our desire to develop a personality assessment system that is maximally flexible for civilian and military applications, we focused on identifying a comprehensive set of nonredundant narrow traits, which, if desired, can be combined to form the Big Five or other broad traits, such as Integrity or Positive Core Self-Evaluations (Judge, 2009). Our literature review revealed that there was no widely agreed upon taxonomy for the narrow (lower-order) personality traits. The majority of existing taxonomies are based either on unverified theoretical assumptions (e.g., the AB5C Circumplex model by Hofstee, De Raad, & Goldberg, 1992), rational judgments (e.g., Hough & Ones, 2002), or researchers' own intuitions about the lower-order structure of a specific Big Five factor (e.g., the NEO Personality Inventory by Costa, McCrae, & Dye, 1991). Although all of these views are interesting, we wanted the TAPAS trait taxonomy to be empirically based to reduce redundancy among the facets, so we reviewed and conducted factor analytic studies using research participants' responses to a diverse array of personality indicators. The first investigation, by Saucier and Ostendorf (1999), examined the structure of a comprehensive set of adjectives describing human behavior. The main assumptions were that all important personality traits have been encoded in the human lexicon and studying the covariation among trait descriptors would thus lead to identification of these traits. Specifically, Saucier and Ostendorf (1999) factor analyzed 312 vectors of responses to 500 personality adjectives and derived a hierarchical taxonomy that consisted of the Big Five broad factors, each of which comprised 3-4 narrow facets.

The second empirical investigation, conducted by the members of our research team together with researchers from the University of Oregon (Professor Lewis Goldberg) and University of Illinois at Urbana-Champaign (Professor Brent Roberts), focused on the universe of available personality scales. The assumption was that most, if not all, important lower-order personality factors have been identified and measured in one form or another by already existing personality inventories, because these inventories have been developed under differing theoretical and research traditions. Specifically, we examined data from a sample of 737 members of the Eugene and Springfield communities in Oregon who, over a period of five years, completed seven major personality inventories: the revised NEO Personality Inventory (NEO-PI-R; Costa, McCrae, & Dye, 1991), the Sixteen Personality Factor Questionnaire (16PF; Conn & Rieke, 1994), California Personality Inventory (CPI; Gough, 1987), the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982), the Jackson Personality Inventory – Revised (JPI-R; Jackson, 1994), the Hogan Personality Inventory (HPI; Hogan & Hogan, 1992), and the Abridged Big Five-Dimensional Circumplex (AB5C) scales from the International Personality Item Pool (Goldberg, 1997) By factor analyzing responses to lower order scales contained in these seven personality measures (e.g., the homogeneous item composites of the HPI), members of our research team were able to derive a lower-order taxonomy for Conscientiousness involving six facets (Roberts, Chernyshenko, Stark, & Goldberg, 2005)**,** and similar analyses were conducted to develop lower-order taxonomies for Emotional Stability, Agreeableness, Openness to Experience, and Extraversion. The result was 22 personality facets that could be located within the Big Five.

The TAPAS Representation of Conscientiousness

The lexical study by Saucier and Ostendorf (1999) found four narrow Conscientiousness facets: Order, Decisiveness, Reliability, and Industriousness. Each facet exhibited a .7-.8 loading on the latent factor, as well as correlations of .65-.85 with a questionnaire-based Conscientiousness scale. Both sets of results were interpreted as evidence that the four facets belong to the Conscientiousness domain. The Order facet was marked by such adjectives as organized, orderly, neat, and meticulous (positive pole of the trait) and disorganized, disorderly, sloppy, and unsystematic (negative pole of the trait). The adjectives for the Decisiveness facet were decisive, firm, consistent, and steady vs. indecisive, inconsistent, scatterbrained, and illogical. The Reliability facet, which is also often called Responsibility or Dependability, was marked by such adjectives as reliable, dependable, responsible, prompt, punctual, and respectful vs. undependable, and unreliable. Finally, the Industriousness facet had on its positive end ambitious, industrious, and purposeful, while on its negative end it had aimless, negligent, and lazy.

The questionnaire-based study of the Conscientiousness domain, published in *Personnel Psychology* (Roberts, Chernyshenko, Stark, & Goldberg, 2005), involved factor analysis of responses to 36 lower order scales identified as measuring various aspects of the Conscientiousness domain. A six-factor solution was found to be most appropriate for describing the observed scale correlation matrix. The first facet, labeled Achievement, was defined by four scales from the NEO-PI (Competence, Achievement-striving, Self Discipline, and Dutifulness) and by four scales from the AB5C (Organization, Purposefulness, Efficiency, and Rationality). Individuals with high scores on this factor would be described as hard working, ambitious, confident, and resourceful.

The second facet, labeled Order, was defined by high loadings on three AB5C scales (Orderliness, Conscientiousness, and Perfectionism), the Order scale from the NEO-PI, the Organization scale from the JPI, and the Perfectionism scale from the 16PF. All of these scales emphasize the ability to plan and organize tasks and activities.

The third facet, Self Control, was defined by the Self-control scale of the MPQ, the Cautiousness scale of the AB5C, the Deliberation scale of the NEO-PI, and the Impulse Control scale of the HPI. Individuals with high scores on Self Control tend to be cautious, patient, levelheaded, and able to delay gratification. In contrast, individuals with low Self Control scores tend to be impulsive, spontaneous, easily distracted, and careless.

The fourth facet, labeled Responsibility, was defined primarily by the Responsibility, Achievement via Conformance, and Socialization scales from the CPI, the Responsibility scale from the JPI, and the Avoids Trouble scale of the HPI. Individuals with high Responsibility scores like to be of service to others, frequently contribute their time and money to community projects, and tend to be cooperative and dependable.

The fifth facet, labeled Non-Delinquency, was defined by the Traditionalism scales of the MPQ and JPI, as well as by the Rule-consciousness scale of the 16PF. People with high scores on Non-Delinquency tend to comply with current rules, customs, norms, and expectations; they dislike change and do not challenge authority.

The sixth facet, Virtue, represents a constellation of beliefs and behaviors associated with adherence to standards of honesty, morality, and “good Samaritan” behavior. The scales that defined this factor were the Good Impression, Self-control and Well-being scales from the CPI, and the Virtuous and Moralistic scales from the HPI. Ostensibly, individuals who score high on this dimension have a tendency to act in accordance with accepted rules of good, or moral, behavior and strive to be a moral exemplar.

The rotated exploratory factor analysis (EFA) pattern matrix for the 32 scales and the six-facet Conscientiousness structure is presented Table 2.1. In the table, each column represents a facet. Values in each row indicate the loadings of a particular personality scale on each facet with primary loadings indicated in bold font.

Table 2.1. The 6-Facet EFA Solution for Conscientiousness

Scale Name	Conscientiousness Facets					
	Achievement	Order	Self Control	Responsibility	Non-delinquency	Virtue
NEO Competence	**.88**	-.28	.14	.10	-.01	-.09
NEO Achievement Striv	**.76**	.02	-.12	.10	.09	-.18
AB5C Organization	**.75**	.11	.05	.11	-.10	-.17
AB5C Purposefulness	**.67**	.18	-.04	-.02	-.11	.24
NEO Self-Discipline	**.65**	.22	-.11	-.03	-.02	.16
AB5C Efficiency	**.63**	.36	-.19	-.03	-.07	.21
AB5C Rationality	**.50**	.16	.12	-.28	.16	-.01
NEO Dutifulness	**.49**	-.05	.14	-.02	.26	.09
AB5C Dutifulness	.22	.12	.17	.17	.14	.11
AB5C Orderliness	-.08	**.86**	.01	.03	.03	.03
NEO Order	.12	**.78**	.02	-.01	-.05	-.09
16PF Perfectionism	.02	**.72**	.08	-.03	.07	-.03
JPI Organization	.16	**.62**	.09	.05	-.02	-.05
AB5C Conscientiousnes	.35	**.61**	.01	.00	-.10	.08
AB5C Perfectionism	.17	**.60**	.06	-.02	.09	-.37
HPI Mastery	.16	.29	.02	-.01	.22	-.01
AB5C Cautiousness	.09	.02	**.75**	-.07	-.04	.02
NEO Deliberation	.37	-.23	**.72**	-.11	-.01	.06
MPQ Self-Control	.10	.16	**.69**	.01	-.05	-.03
HPI Impulse Control	-.25	.14	**.63**	.04	.01	.19
HPI Not Spontaneous	-.05	.05	**.45**	.10	-.06	-.08
MPQ Harm Avoidance	-.24	.18	.33	.15	.06	-.01
CPI Responsibility	.09	-.02	-.05	**.90**	-.06	-.02
CPI Achievement via Conformance	.35	-.03	.04	**.71**	-.03	-.07
CPI Socialization	.03	-.01	.10	**.52**	.11	.17
JPI Responsibility	.00	-.04	-.09	**.51**	.38	.06
HPI Avoids Trouble	-.20	.11	.13	**.40**	-.04	.19
HPI Not Autonomous	-.18	.11	.09	.34	.05	-.27
MPQ Traditionalism	.01	.01	-.04	-.08	**.91**	.04
JPI Traditionalism	-.04	.00	-.04	.06	**.85**	.10
16PF Rule Consciousne	-.02	.12	.04	.22	**.63**	.03
CPI Good Impression	.03	-.06	-.01	.14	.04	**.80**
CPI Self-Control	-.21	-.01	.25	.12	-.03	**.78**
CPI Well-Being	.21	-.13	-.12	.33	-.17	**.48**
HPI Moralistic	.09	.01	-.12	-.20	.33	**.47**
HPI Virtuous	-.02	-.13	.03	-.03	.13	**.44**

Note. N=734.

Note that the six-facet representation stemming from the questionnaire-based study subsumes the four-facet structure derived from the lexical study. The two taxonomies shared

Order, Achievement, and Responsibility (Reliability). The lexical factor, Decisiveness was essentially absorbed by Industriousness in the questionnaire-based study because many current scales measuring achievement contain decisiveness items, such as “I tend to finish tasks that I’ve started” or “I don’t stop working on a project until I finish it.” Although the Non-Delinquency, Self Control, and Virtue factors were not identified in Saucier and Ostendorf’s lexical study, a more recent study by Roberts, Bogg, Walton, Chernyshenko, and Stark (2004), which included an expanded set of adjectives describing Conscientiousness, clearly found these facets. Consequently, given these replicable empirical results, we decided to adopt the six-facet structure for TAPAS.

To help with the interpretation of the six facets and to see how these lower-order factors related to the broad Conscientiousness factor, we performed a hierarchical analysis by computing factor scores for the extractions with one through six factors and related these factor scores across solutions. The results of this procedure, depicted in Figure 2.1, yielded a “top-down” representation of the proposed structure; the coefficients in the figure are the correlations between the factor scores from the factors at each level with those at levels above and below it. As can be seen in Figure 2.1, in the two-factor solution, the thirty-six scales first broke into what we call proactive and inhibitive aspects of Conscientiousness. In the three-factor solution, a subcomponent of proactive aspects became an achievement factor, and the remaining proactive aspects of conscientiousness combined with the inhibitive aspects to create an integrity factor consisting of scales measuring social responsibility and honesty and a rule-orientation factor containing scales measuring traditionalism and impulse control. In the four factor solution the rule-orientation factor was further divided into self-control and traditionalism factors. Moving to the fifth level, the integrity factor was split into responsibility and virtue factors. Finally, the six-factor solution resulted from the achievement factor splitting into industriousness and order, yielding the final set of TAPAS Conscientiousness facets: *Achievement, Order, Responsibility, Virtue, Self Control, and Non-Delinquency*.

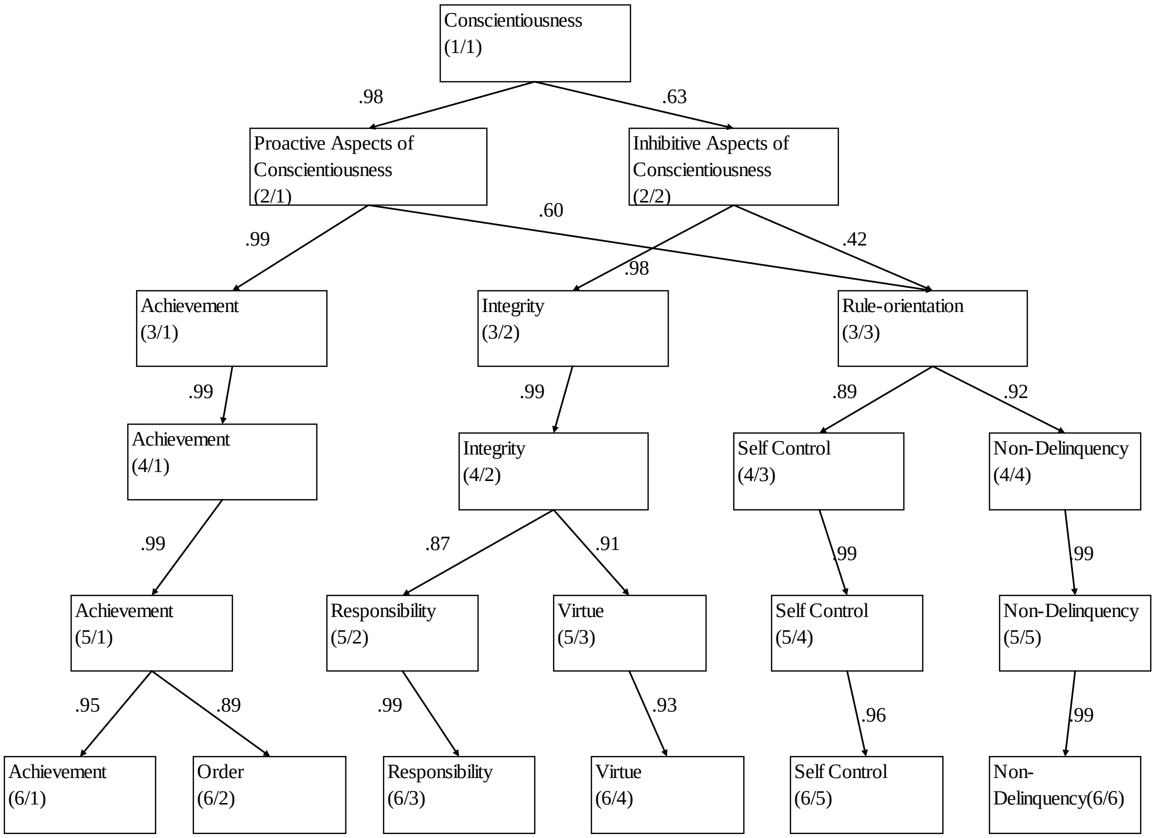

Figure 2.1. The Hierarchical Structure of Conscientiousness.

A more thorough examination of questionnaire results revealed a very important outcome that was directly relevant to TAPAS development: *None of the available personality inventories measured all six lower-order facets of Conscientiousness; most assessed just two or three.* For example, five of six NEO-PI scales and eight of nine AB5C scales loaded primarily on the Industriousness and Order facets, and no scales from those two inventories loaded highly on the Responsibility and Virtue facets. Because we wanted TAPAS to provide the most complete assessment of the underlying structure of the Big Five factors, we retained all six. The intercorrelations of the TAPAS facets assessing Conscientiousness are shown in Table 2.2.

Table 2.2. Correlations among the TAPAS Facets of Conscientiousness

Facet	Achievement	Order	Self Control	Responsibility	Non-Delinquency	Virtue
Achievement	1.00					
Order	.71	1.00				
Self Control	.47	.58	1.00			
Responsibility	.27	.15	.43	1.00		
Non- Delinquency	.25	.51	.47	.17	1.00	
Virtue	.51	.30	.52	.62	.23	1.00

As can be seen in the table above, the intercorrelations are all positive, but not excessively high (the average was .41), indicating that large amounts of facet-specific variance are present. This is important because it suggests that forming composites from TAPAS facet scores may provide incremental validity in applied settings. Developing pools of statements for each facet and combining scores after the fact is also advantageous from a psychometric standpoint, because it facilitates calibration of the statement pools using unidimensional IRT models, such as the GGUM, as was discussed in Chapter 1.

The TAPAS Representation of Openness to Experience

Openness to Experience is regarded as one of the key personality variables for explaining and understanding behavior of individuals in settings characterized by high levels of uncertainty and change (Hough, 2003). However, because of a divergence of views among researchers about the precise structure of this broad construct, the use of Openness measures in applied research has been limited. Even at the broadest level, there is a disagreement whether Openness to Experience should be viewed solely as intellect (ability to efficiently process information or create new ideas) or whether it should also include other, less intellectualized behaviors, such as tolerance, fantasy, and interest in artistic experiences (Digman, 1990; Goldberg, 1993; McCrae, 1996). Empirical studies of the sort described here can provide grounds for clarifying the nature of this nebulous construct.

A lexical study by Saucier and Ostendorf (1999) involved factor analysis of responses to adjectives believed to be associated with Openness. Three facets were found: Imagination, Intellect, and Perceptiveness. The Imagination facet had the highest loading on the Openness factor (.87) and was marked by adjectives such as creative, inventive, clever, innovative (positive end of the trait continuum), uncreative, and unimaginative (negative end of the trait continuum). The adjectives for the Intellect facet were intelligent, analytical, and knowledgeable vs. nonintellectual and unreflective. The Perceptiveness facet, which had the lowest loading on Openness (.64), was marked by adjectives such as perceptive, insightful, and foresighted vs. imperceptive, unobservant, and shortsighted. From the content inspection of each facet, it was apparent that the adjective analysis adopted a narrower view of Openness to Experience; namely, it focused on behaviors associated with the domain of intellectual functioning.

Our questionnaire-based study, on the other hand, revealed a much broader configuration of the Openness construct. The study involved an exploratory factor analysis of scores on 34 scales from seven widely used personality inventories and identified a stable, six-facet structure. The composition of the first three facets essentially mirrored the lexical study and contained scales that emphasized cognitive competence, ingenuity, and curiosity. The first facet was named Intellectual Efficiency and was defined by three scales from the HPI (Education, Good Memory, and Reading), the Intellectual Efficiency scale from the CPI, and Intellect and Quickness scales from the AB5C. Individuals with high scores on this facet are able to process information quickly and would be described by others as knowledgeable, astute, and intellectual. Hence, this facet was largely equivalent to the Intellect facet found in the lexical study.

The second facet, labeled Ingenuity**,** was defined by high loadings on the Ingenuity and Competence scales from the AB5C, Generate Ideas from the HPI, and the Innovation scale from the JPI. A prototypical individual scoring high on the Ingenuity facet is an inventor, a person who constantly strives to make improvements to the existing information or products. The Ingenuity facet found in our questionnaire study shared many features with the Imagination facet found in the lexical study.

The third Openness facet was named Scientific Curiosity or Curiosity for short**,** because it was defined by the Curiosity, Science Ability, and Thrill Seeking scales from the HPI, and the Sensitivity scale of the 16PF. The majority of items contained in these scales described behaviors directed toward understanding how the world around us "works." Individuals with high scores on this facet would be characterized as inquisitive and perceptive, they read popular science/mechanics magazines, and, at least at some point of their lives, they have conducted physics or chemistry experiments or have disassembled and reassembled a piece of machinery or an electrical appliance.

The other three facets, however, were composed of scales related to less intellectualized behaviors, such as traveling abroad, learning new cultures and languages, collecting art and/or participating in artistic activities, and striving toward self understanding. Factor analysis organized these scales into Aesthetics, Tolerance, and Depth. The Aesthetics facet was defined by the Aesthetics and Feelings scales from the NEO-PI, Reflection and Imagination scales from the AB5C, the Absorption scale from the MPQ, the Culture scale from the HPI, and the Breadth of Interests scale from the JPI. Most items in these scales were concerned with artistic/aesthetic experiences. Individuals scoring high on the Aesthetics facet genuinely enjoy acquiring, participating, or creating various forms of artistic, musical, or architectural outputs. Unlike individuals high on the Curiosity facet, they are not necessarily interested in understanding how or why things they enjoy were created; instead, they are more interested in the experiential component of the behavior.

The fifth Openness to Experience facet was labeled Tolerance, because it was defined by the Flexibility and Psychological Mindedness scales from the CPI, the Values scale from the NEO-PI, and the Tolerance scale from the JPI. As is evident from the name, this facet deals with behavior toward strangers and, more generally, novel stimuli. Individuals scoring high on Tolerance are comfortable with people speaking a foreign language around them or expressing a different viewpoint. They are interested in learning about different cultures and they often attend cultural events or meet and befriend people from around the world. When given a chance to travel, their intent is to immerse themselves into new customs and traditions, rather than just to enjoy the scenery.

The final Openness facet, Depth, was defined by the Abstractedness scale from the 16PF, the Introspection scale from the AB5C, and the Complexity scale from the JPI. All of these scales measure behaviors aimed at understanding one's self and/or facilitating self- improvement and self-actualization. Examples of such behaviors include reflection, meditation, introspection, attending personal growth seminars, and seeking spiritual enlightenment. The complete rotated pattern matrix for the 36 scales and the six-facet Openness to Experience structure is presented in Table 2.3.

Table 2.3. The 6-Facet EFA Solution for Openness to Experience

Scale Name	Openness To Experience Facets					
	Intellectual Efficiency	Ingenuity	Curiosity	Aesthetics	Tolerance	Depth
HPI Education	**.69**	-.14	-.05	-.17	-.01	.09
HPI Good Memory	**.60**	.17	-.12	-.01	-.21	.00
HPI Reading	**.56**	-.04	-.37	-.01	.09	.16
CPI Intellectual Efficiency	**.56**	.14	.02	-.05	.47	-.33
AB5C Intellect	**.54**	.25	-.07	.04	.02	.26
AB5C Quickness	.51	.47	.05	-.03	-.09	.07
NEO Ideas	.38	.02	.30	.18	-.04	.33
HPI Intellectual Games	.33	-.16	.22	.12	-.07	-.13
AB5C Ingenuity	-.11	**1.01**	-.07	.02	.09	-.08
HPI Generates Ideas	-.05	**.72**	-.04	-.02	.15	-.02
AB5C Competence	.35	**.66**	.06	-.03	-.24	-.08
JPI Innovation	-.22	**.54**	.15	.15	.17	.18
16PF Sensitivity	.31	-.14	**-.72**	.44	.07	.09
HPI Curiosity	-.09	.03	**.65**	.12	-.16	.05
HPI Science Ability	.08	.04	**.63**	.08	.02	.03
HPI Thrill Seeking	-.11	-.10	**.46**	.00	.04	.17
HPI Math Ability	.21	-.06	**.37**	-.24	-.06	-.01
NEO Aesthetics	-.03	-.06	.03	**.99**	.05	-.15
AB5C Reflection	-.01	.05	-.13	**.96**	-.09	-.19
MPQ Absorption	-.23	-.03	.12	**.72**	-.18	.24
NEO Feelings	-.11	.27	-.19	**.53**	-.13	.12
HPI Culture	.25	-.31	.16	**.52**	.18	-.01
AB5C Imagination	-.08	.07	-.01	**.52**	.29	.18
JPI Breadth	.17	-.07	.22	**.46**	.20	-.04
NEO Actions	-.11	.20	-.01	.36	.31	-.09
CPI Flexibility	-.05	-.10	-.11	-.20	**.88**	.11
NEO Values	-.08	.09	-.08	.00	**.58**	.05
CPI Psychol. Mindedness	.40	.11	.11	-.05	**.50**	-.32
JPI Tolerant	-.01	.06	-.04	.16	**.46**	-.08
NEO Fantasy	-.20	.14	-.02	.11	**.41**	.32
16PF Openness to Change	.01	.11	.07	.14	**.35**	.28
16PF Abstractness	-.26	-.09	.05	-.14	.33	**.73**
AB5C Depth	.11	.04	.04	.14	-.23	**.68**
AB5C Introspection	.10	-.03	.03	-.11	-.07	**.64**
JPI Complexity	.27	-.17	-.03	.07	.29	**.46**
AB5C Creativity	.37	.24	.19	-.13	.03	.43

Note. N = 747.

As was noted above, the six-facet structure that emerged from the questionnaire study is consistent with the broad concept of Openness to Experience and, unlike the structure from the lexical study, it includes a number of behavioral clusters only moderately related to intellectual functioning. In this respect, our structure aligns more closely with the views of McCrae and Costa (1997), who argued for the inclusion of aesthetics, feelings, and tolerance (also labeled “values” in their nomenclature). This is not particularly surprising, given that scales from the NEO-PI were included in our questionnaire analysis. However, what was surprising is the degree of interrelationship among the six facets. While it was true that the three “intellectual” and the three “nonintellectual” facets intercorrelated highly within their subgroupings, the correlations across the two subgroupings were also reasonably high (ranging up to .55). These results clearly indicate a hierarchical structure, with a broad Openness factor at the top of a hierarchy splitting into two narrower factors, Creative Intellect and Breadth of Interests/Values, which in turn split into even narrower factors to produce a total of six interpretable facets. The complete hierarchical solution for Openness to Experience is presented in Figure 2.2. As in the Conscientious investigation, this hierarchical representation was established by computing factor scores for extractions involving one to six factors and correlating the factors scores across successive levels in the hierarchy.

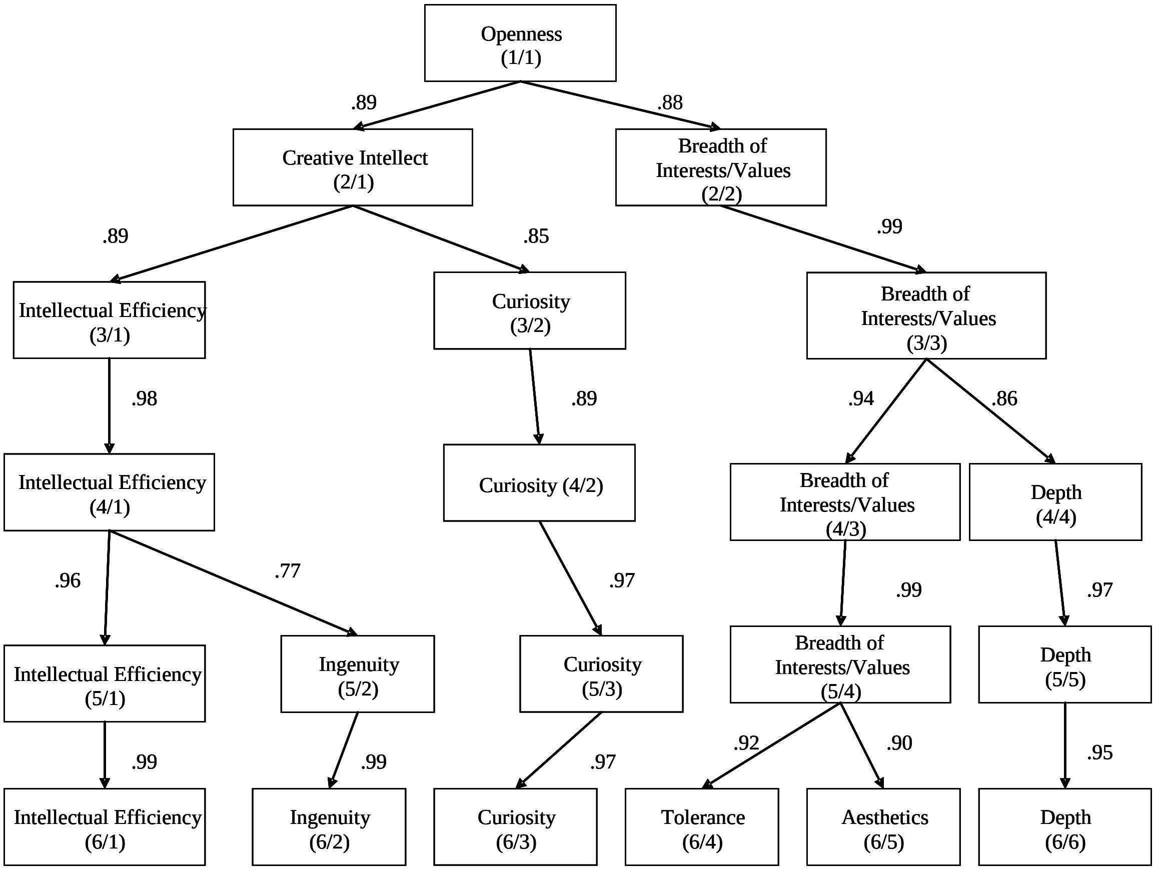

Figure 2.2. The Hierarchical Structure of Openness to Experience.

In sum, combining results from the two factor analytic investigations involving Openness to Experience indicators produced a six-facet solution: *Intellectual Efficiency, Ingenuity,* *Scientific Curiosity (abbreviated Curiosity), Tolerance, Aesthetics, and Depth*. These six facets form the TAPAS representation of Openness, with intercorrelations as shown in Table 2.4. For clarity it is important to note that Intellectual Efficiency is not a measure of cognitive ability. Although Intellectual Efficiency scores correlate about .4 with measures of general cognitive aptitude measures, such as the Armed Forces Qualification Test (AFQT), Intellectual Efficiency focuses on one's interest in acquiring knowledge and the perceived ease or speed with which one make's decisions, rather than the level or accuracy of one's knowledge or the quality of those decisions.

Table 2.4. Correlations between the TAPAS Facets of Openness to Experience

Facet	Aesthetics	Intellectual Efficiency	Tolerance	Ingenuity	Depth	Curiosity
Aesthetics	1.00					
Intellectual Efficiency	.44	1.00				
Tolerance	.67	.55	1.00			
Ingenuity	.44	.54	.51	1.00		
Depth	.69	.36	.52	.58	1.00	
Curiosity	.09	.31	.26	.54	.27	1.00

Developing the TAPAS Lower-Order Representations of Extraversion, Agreeableness, and Emotional Stability

Our initial inspection of the scales measuring lower-order factors of Extraversion, Emotional Stability, and Agreeableness in the seven major questionnaires revealed many inconsistencies. For example, the NEO-PI aligns Warmth with the Big Five factor Extraversion, but Warmth is considered part of Agreeableness in the AB5C. Sometimes even scales having the same or very similar names differed markedly in content. Consider, for example, two scales named Cooperativeness. Whereas the JPI Cooperativeness items focus on an individual's insecurities, the AB5C items tap aspects of hostility, dominance, unrestraint, and self-control. Therefore, despite the intuitive connection between the name Cooperativeness and the Big Five factor Agreeableness, the respective scales actually measure one or more of the *other* Big Five factors.

Inconsistencies such as these made it impossible to identify scales *a priori* that distinctly measured Extraversion, Agreeableness, or Emotional Stability for separate investigations of their lower-order factor structures. Consequently, to provide a basis for such analyses, we pooled the data for the 99 scales associated with any of these three factors and conducted a joint exploratory factor analysis. Figure 2.3 presents the scree plot for this analysis. As can be seen in the figure, the first three factors accounted for almost half of the total variance, with all three having eigenvalues greater than 10. Examination of the Promax rotated pattern matrix for this 3-factor solution revealed 30 scales loading primarily on Extraversion, 30 loading primarily on Emotional Stability, and 21 loading mainly on Agreeableness. The remaining 18 scales either loaded on more than one factor or had uniformly low loadings (below .35) and were thus dropped from further analyses to provide cleaner factor solutions. In sum, once the scales loading primarily on just one Big Five factor were identified, each subset of scales was factor analyzed separately to identify the lower-order structures for Extraversion, Agreeableness, and Emotional Stability needed for TAPAS.

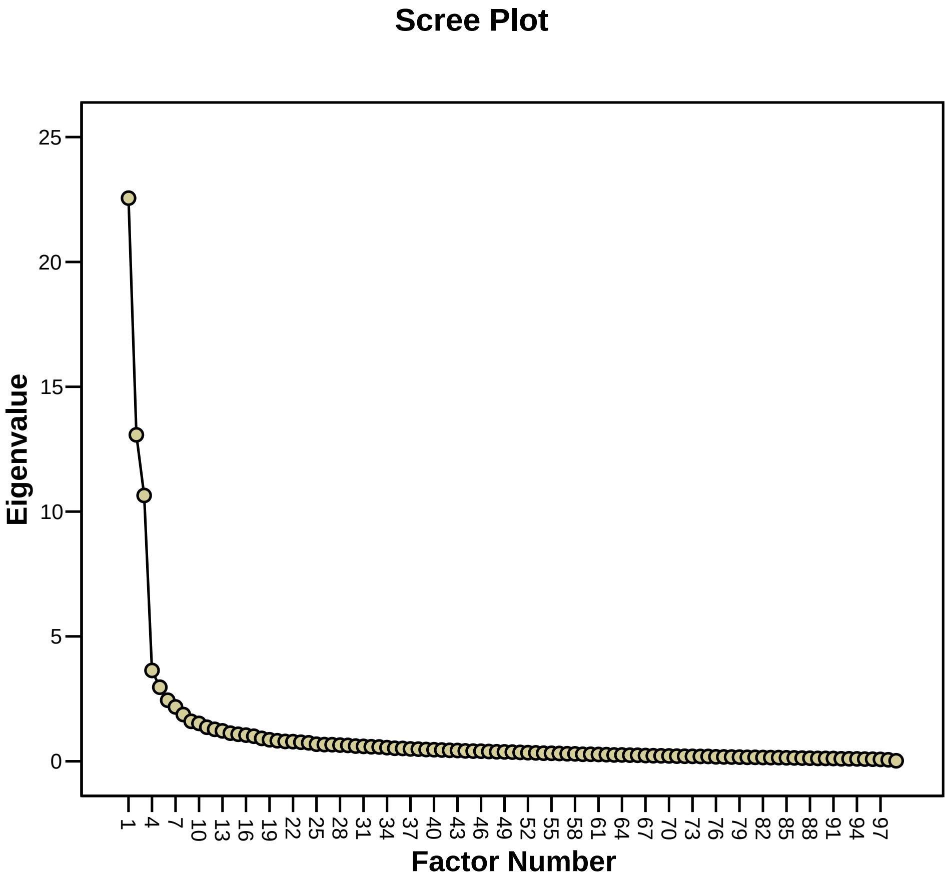

Figure 2.3.* *Scree Plot from the EFA of 99 Scales Belonging to the Extraversion, Agreeableness, and Emotional Stability Domains.

The TAPAS Representation of Extraversion

A majority of researchers would agree that Extraversion includes such behavioral classes as sociability and assertiveness (Hough & Ones, 2002; Mershon & Gorsuch, 1988). Yet there has been much disagreement about what other behaviors should be included. Some authors add an energy or activity level component (Digman, 1990), while others add excitement seeking (McCrae & Costa, 1985; Norman, 1963) or even friendliness and warmth (Cattell, 1973; Goldberg, 1993). This uncertainty is partially due to a lack of agreement among early test developers as to the theoretical links between potential facets. Fortunately, recent research aimed at determining the fundamental features of Extraversion (e.g., Ashton, Lee, & Paunonen, 2002; Lucas, Diener, Grob, Suh, & Shao, 2000) has brought some clarity to this domain. Specifically, recent studies have investigated whether Extraversion should be viewed primarily as 1) a preference for social interactions, 2) a tendency to experience pleasant affect across a variety of rewarding situations (reward sensitivity), or 3) a tendency to engage and enjoy social attention for its own sake. Each view leads to a somewhat different configuration of the broad factor, so determining which one is the most consistent theoretically and empirically is important in the context of taxonomy building.

Recent empirical results by Ashton et al. (2002) support choosing social attention as a fundamental feature of Extraversion. From this viewpoint, a three-facet representation of Extraversion suffices: Affiliation (tendency to engage and enjoy friendly social interactions), Ascendance (tendency to enjoy leadership, dominance, and assertive behaviors), and Venturesomeness (tendency to enjoy exciting social interactions, such as parties). Although this three-facet structure is compelling, its validity must ultimately be supported by comprehensive empirical studies, similar to those appearing in this report. Hence, we discuss our results from the lexical and questionnaire studies in the context of this taxonomy.

The lexical study by Saucier and Ostendorf (1999) identified four Extraversion facets: Sociability, Unrestraint, Assertiveness, and Activity/Adventurousness. The Sociability facet had the second highest loading on the Extraversion factor (.81) and was marked by adjectives such as sociable, cheerful, and merry (positive end), and unsociable, withdrawn, and uncommunicative (negative end). The Unrestraint facet had the highest Extraversion loading (.85) and contained adjectives such as talkative, verbal, aggressive vs. quiet, nontalkative, and reserved. The Assertiveness facet, which had the lowest loading on the Extraversion factor (.76), was marked by such adjectives as assertive, direct, and bold vs. weak, submissive, and helpless. The last facet, Activity, included, at the positive pole, adjectives such as active, competitive, adventurous, and, at the negative pole, unadventurous, uncompetitive, and unenergetic. Comparing this four-facet representation to Ashton et al.'s (2002) taxonomy revealed that the Unrestraint facet most closely aligned with the Ascendance facet (both involve leadership and assertive activities), the Sociability facet most closely resembled the Affiliation facet, and the combination of at least some aspects of the Unrestraint and Activity/Adventurousness facets in the Saucier and Ostendorf's taxonomy could be viewed as the Venturesomeness facet in the other conceptualization. The main difference between the two views appears to be in the activity/energy component, which is clearly present in the Saucier and Ostendorf's study, but apparently dispersed among Ascendance and Venturesomeness in Ashton et al.'s taxonomy.

Our questionnaire-based investigation involved factor analyzing responses to 30 Extraversion-related scales (see the section immediately above for an explanation of how these scales were selected) and resulted in a three-factor solution. The rotated factor loading matrix from an EFA is shown in Table 2.5.

As can be seen from Table 2.5, three facets of Extraversion were identified. The first facet could have been named Dominance/Energy, because its best markers were the Leadership scales of the HPI and AB5C, the Assertiveness scales from the NEO-PI and AB5C, the Social Potency scale from the MPQ, and the Dominance scale from the CPI. Individuals with high scores on Dominance/Energy are assertive, dominant, and would be described by their peers as "leaders." However, because the first facet also contained two scales pertaining to one's level of activity (the Energy scale from the JPI and the Activity scale from the NEO-PI) and because the activity themes formed a separate facet in the lexical study, we chose the name

Dominance/Activity instead. The important difference between the EFA findings in the lexical and questionnaire studies is that the lexical study included many markers for activity, whereas the questionnaire study involved just two. Hence, the apparent lack of an activity facet in the questionnaire study might be explained by sampling; there were simply not enough activity markers in the questionnaire study to produce an interpretable factor. Ultimately, whether Dominance/Activity should be split into two facets or kept as one is an empirical issue, which requires further study. For TAPAS development, we chose to separate the dominance and activity components for two reasons. First, it can be argued that dominant individuals are generally active, but active people are not necessarily assertive or dominant. Second, splitting the components provides consistency with other personality scales used in military contexts, such as the AIM. In the AIM, dominance behaviors are captured by a scale called Leadership and activity behaviors are tapped by Physical Conditioning.

The second facet extracted in our questionnaire-based investigation was labeled Sociability. It was defined by high loadings on the Sociability, Capacity for Status, Empathy, and Social Presence scales from the CPI, the Poise scale from the AB5C, the No Social Anxiety scale from the HPI, and the Social Boldness scale from the 16PF. All of these scales describe individuals who are interested in social interactions, on the positive end, or those who prefer to avoid them at the negative end. Note that both the lexical study and Ashton et al.'s studies also found Sociability to be one of the main features of the Extraversion dimension.

The third facet, labeled Attention Seeking, was marked by the Liveliness scale from the 16PF, the Excitement Seeking scale from the NEO-PI, the Self-Disclosure scale from the AB5C, as well as four HPI scales: Likes Crowds, Likes Parties, Exhibitionistic and Entertaining. This facet was very similar to the Venturesomeness facet found in Ashton et al.'s study and the Attention Seeking facet found in the lexical study. In essence, behaviors from this facet target the excitement-seeking component of social interactions, rather than just one's participation in social exchanges (the latter is more indicative of the Sociability facet). Individuals scoring high on the Attention Seeking facet engage in behaviors that attract a lot of social attention; they are loud, loquacious, entertaining, and even boastful. Note, however, these behaviors do not necessarily indicate a lack of self-control (one of the facets of Conscientiousness), but, rather, intentional behaviors that bring substantial enjoyment and satisfaction.

Table 2.5. The 3-Facet EFA Solution for Extraversion

Scale Name	Extraversion Facets		
	Dominance/Activity	Sociability	Attention Seeking
HPI Leadership	**.87**	-.09	-.01
NEO Assertiveness	**.84**	.05	-.02
MPQ Social Potency	**.76**	-.04	.24
AB5C Assertiveness	**.74**	.10	-.15
CPI Dominance	**.69**	.41	-.21
AB5C Provocativeness	**.59**	-.21	.37
AB5C Leadership	**.58**	.34	.01
HPI Competitive	**.52**	.04	-.07
NEO Activity	**.52**	.00	.07
NEO Modesty	**-.38**	-.09	-.07
JPI Energy	**.37**	.23	-.13
HPI Experience Seeking	.21	.15	.21
CPI Capacity for Status	-.08	**.89**	-.05
CPI Sociability	.03	**.84**	.08
AB5C Poise	.02	**.80**	-.04
CPI Empathy	-.07	**.76**	.08
CPI Social Presence	-.08	**.72**	.26
HPI No Social Anxiety	.24	**.69**	-.21
16PF Social Boldness	.18	**.63**	.06
CPI Self-Acceptance	.37	**.57**	-.02
JPI Confidence	.39	**.49**	.10
AB5C Gregariousness	.09	.43	.41
16PF Liveliness	-.21	.14	**.80**
NEO Excitement Seeking	.11	-.20	**.62**
AB5C Self-Disclosure	-.15	.22	**.61**
HPI Likes Crowds	-.10	.04	**.54**
HPI Likes Parties	-.10	.28	**.47**
HPI Entertaining	.14	.07	**.46**
HPI Exhibitionistic	.25	.02	**.46**
NEO Straightforwardness	-.39	.36	-.41

Note. N = 747.

In summary, three correlated facets of Extraversion were identified in our questionnaire study by factor analysis of responses to 30 Extraversion scales. However, for consistency with the lexical findings of Saucier and Ostendorf (1999) and the U.S. Army AIM questionnaire, we adopted a four-facet representation of Extraversion for TAPAS.

To produce a hierarchical structure similar to what was developed for Conscientiousness and Openness to Experience, we started with the EFA results from the questionnaire study and added a step. Specifically, we produced the first three levels of the hierarchy by computing factor scores for the EFA extractions involving one through three factors and correlating the scores from successive solutions. To produce the fourth and final level of the hierarchy, which shows the split of the dominance and activity components, we computed z-scores for all scales having primary loadings on the Dominance/Activity facet (third level of hierarchy) and then formed two unit-weighted composites consisting of dominance-related scales (i.e., leadership, assertiveness, etc.) and activity-related scales (i.e., activity and energy). These were then correlated with factors from other levels in the hierarchy to produce the complete hierarchical solution shown in Figure 2.4.

In Figure 2.4, it can be seen that the broad factor of Extraversion first split into Sociability and Dominance/Activity components. Then, at the next level of the hierarchy, the Sociability facet divided into Sociability (participating in social interactions) and Attention Seeking (seeking excitement from social interactions). Finally, at the fourth level, the Dominance/Activity facet split into its respective components to produce the final four facets of Extraversion included in TAPAS: *Attention Seeking, Sociability, Dominance, and Activity*. Importantly, for continuity with terminology in other military research, we subsequently renamed the activity facet *Physical Conditioning*. The correlations of these four facets are presented in Table 2.6.

Table 2.6. Correlations between the TAPAS Facets of Extraversion

Facet	Dominance	Activity	Sociability	Attention Seeking
Dominance	1.00			
Activity	.51	1.00		
Sociability	.70	.46	1.00	
Attention Seeking	.55	.31	.58	1.00

Note. Dominance and Activity facet scores were unit-weighted composites of scales from the Dominance/Activity factor found in the questionnaire-based study. For continuity with terminology in other military research, Activity was renamed *Physical Conditioning*.

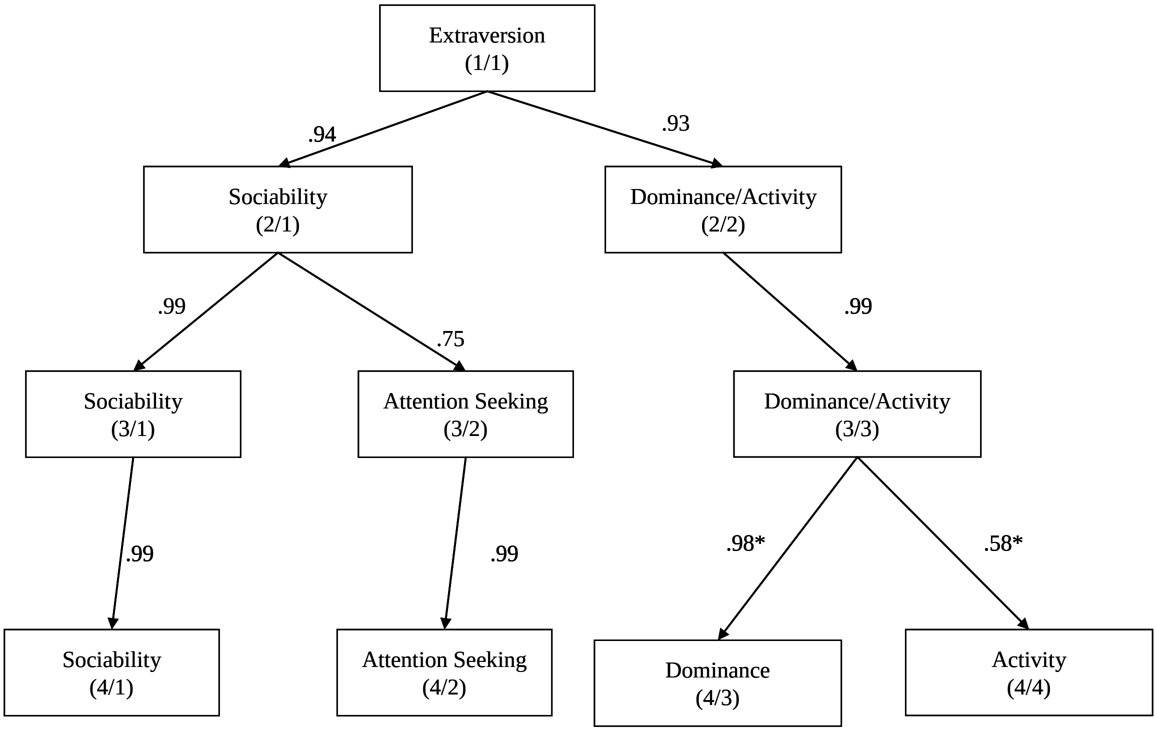

Figure 2.4. The Hierarchical Structure of Extraversion.

Note. Dominance and Activity facet scores were unit-weighted composites of scales from the Dominance/Activity factor found in the questionnaire-based study. For continuity with terminology in other military research, Activity was renamed *Physical Conditioning*.

The TAPAS Representation of Agreeableness

The lexical investigation by Saucier and Ostendorf (1999) found four Agreeableness facets: Warmth-Affectionate, Gentleness, Generosity, and Modesty. Each facet exhibited a .70 - .84 loading on the broad factor and had a correlation of .50 to .82 with a questionnaire-based Agreeableness scale. Both sets of results were interpreted as evidence that the four facets belonged to the Agreeableness domain. The Warmth-Affectionate facet was marked by adjectives such as warm, affectionate, sensitive, and compassionate (positive pole of the trait continuum) and cold, unsympathetic, and insensitive (negative pole of the trait continuum). The adjectives for the Gentleness facet were agreeable, cordial, and amiable vs. antagonistic, rough, and combative. The Generosity facet was marked by adjectives such as charitable, helpful, and generous vs. greedy, stingy, and selfish. Finally, the Modesty facet had modest and humble on its positive end and conceited, snobbish, and egocentric on its negative end.

The questionnaire examination of Agreeableness involved factor analysis of responses to 21 scales identified as part of this domain. We found that a three-factor solution was most appropriate to describe the observed correlation matrix. We named the first facet Cooperation. It was defined by the Pleasantness, Nurturance, and Morality scales from the AB5C, the

Altruism and Trust scales from the NEO-PI, and the Easy to Live With scale from the HPI. Individuals scoring high on Cooperation are trusting, cordial, cooperative, uncritical, kind, and easy to live with, while those scoring low are skeptical, suspicious, and argumentative.

The second facet was named Consideration, because it was defined by three Warmth scales from the AB5C, NEO-PI, and 16PF, the Social Closeness scale from the MPQ and the Positive Emotions scale from the NEO-PI. Individuals scoring high on Consideration are considerate, affectionate, and positive toward others. Unlike extraverts, however, who actively seek social attention, individuals with high Consideration scores may be quite passive socially; they are simply "there for you, whenever needed." Such individuals are confidants and natural, if untrained, psychotherapists.

The last facet found in the questionnaire investigation was named Selflessness. It was marked by the Femininity/Masculinity scale from the CPI, the Sensitivity scale from the 16PF, and four AB5C scales: Sympathy, Tenderness, Understanding, and Empathy. Unlike behaviors associated with the Consideration facet, where the main theme was unconditional positive regard for others, behaviors from the Selflessness facet are more active such as helping and doing things for others, giving to charity and volunteering for community improvement. Individuals scoring high on this facet are generous with their time and resources, sympathetic, and think of others first. Individuals scoring low are selfish, greedy, and even snobbish.

In summary, our questionnaire investigation identified three facets of Agreeableness: *Consideration, Selflessness, and Cooperation*. Comparison of these findings to the lexical representation of Agreeableness revealed that the Selflessness facet was essentially the Generosity facet found by Saucier and Ostendorf, the Consideration facet was nearly identical to the Warmth-Affectionate facet, and the Cooperation facet was a combination of the Gentleness and Modesty facets.

The rotated pattern matrix for the 21 scales of the three facet EFA solution for Agreeableness is presented in Table 2.7. As before, bolded values indicate the primary loadings of the scales on the respective facets. The hierarchical structure of Agreeableness is presented in Figure 2.5, followed by the facet intercorrelation matrix for the third level of the hierarchy in Table 2.8.

Table 2.7. The 3-Facet EFA Solution for Agreeableness

Scale Name	Agreeableness Facets		
	Cooperation	Consideration	Selflessness
AB5C Pleasantness	**.93**	-.07	-.11
AB5C Nurturance	**.74**	-.25	.29
AB5C Morality	**.67**	-.23	.10
NEO Altruism	**.65**	.16	-.01
NEO Trust	**.58**	.31	-.31
HPI Easy to Live With	**.55**	.12	-.22
MPQ Social Closeness	-.15	**.79**	.02
NEO Warmth	.21	**.76**	-.10
NEO Positive Emotions	.14	**.72**	-.21
16PF Warmth	-.21	**.65**	.30
16PF Self Reliance	.14	**-.59**	-.01
AB5C Warmth	.29	**.46**	.29
HPI Caring	.07	**.36**	.17
CPI Femininity/Masculinity	-.10	-.19	**.77**
16PF Sensitivity	-.25	.05	**.73**
AB5C Sympathy	.20	.20	**.57**
AB5C Tenderness	-.07	.31	**.54**
AB5C Understanding	.38	.05	**.50**
AB5C Empathy	.25	.13	**.40**
NEO Tender-Mindedness	.31	-.07	.35
HPI Sensitive	.23	-.03	.26

Note. N = 747.

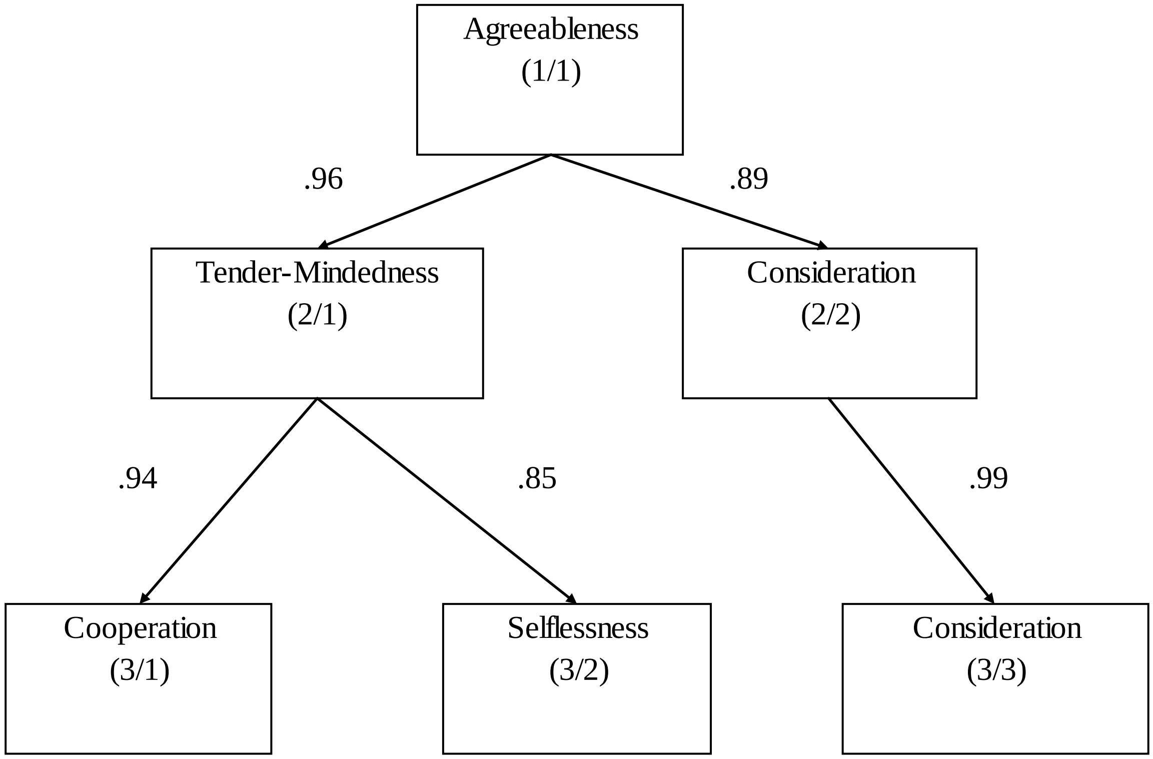

Figure 2.5. The Hierarchical Structure of Agreeableness.

Table 2.8. Correlations between the TAPAS Facets of Agreeableness

Facet	Cooperation	Consideration	Selflessness
Cooperation	1.00		
Consideration	.68	1.00	
Selflessness	.62	.62	1.00

The TAPAS Representation of Emotional Stability

The lexical investigation by Saucier and Ostendorf (1999) found three facets of Emotional Stability: Irritability, Insecurity, and Emotionality. The Irritability facet was marked by adjectives such as irritable and moody (negative end of the Emotional Stability) vs. undemanding and uncritical (positive end of the trait continuum). The adjectives for the Insecurity facet were insecure, unstable, nervous vs. relaxed and unenvious. The Emotionality facet was marked by adjectives such as emotional, anxious, fidgety, and excitable vs. unemotional and unexcitable. Behaviors in all these facets deal with some form of emotional instability/excitability on the negative end and imperturbability/placidity at the positive end.

Our questionnaire examination of the Emotional Stability domain involved factor analysis of responses to 30 scales identified as measuring various aspects of the domain. We found three relatively highly correlated facets: Optimism, Adjustment, and Even Tempered (see the rotated pattern matrix in Table 2.9). The first facet was named Optimism, because it was marked by the No Depression scale from the HPI and the Depression scale from the NEO-PI. In addition, a number of scales describing one's happiness and well-being loaded on this facet (i.e., the Moderation and Happiness scales from the AB5C, the Well-Being scale from the CPI, the Emotional Stability scale from the 16PF, and the No Guilt and Identity scales from the HPI). All these scales try to assess an individual's general emotional tone. The continuum here is joy, well-being, and positive outlook on one end to negative outlook, depressed mood, hopelessness, and despair on the other.

The second facet identified in our questionnaire investigation was Adjustment. It was defined by three Anxiety scales from the JPI, NEO-PI, and HPI, the Apprehension scale from 16PF and the Stress Reaction scale from the MPQ. In addition, the Cooperativeness scale from the JPI showed a very high loading on the Apprehension facet, because it contained a number of items related to one's sensitivity toward the opinions of others. All of these scales describe behaviors associated with various degrees of insecurity and anxiety. Individuals scoring low on the Adjustment facet are high strung, self-conscious and apprehensive in most contexts. This facet essentially mirrors the Insecurity facet found in the lexical investigation, but is scored in the opposite direction.

The third facet, Even Tempered, was defined by the Calmness scale from the AB5C, the Hostility scale from the NEO-PI, the Even Tempered and Empathy scales from the HPI, and the Stability and Tranquility scales from the AB5C. Persons scoring low on this facet tend to experience a range of emotions including irritability, anger, hostility, or even aggression. In contrast, those scoring high on the Even Tempered facet tend to be calm, even-tempered, and stable, even when threatened. This facet most closely resembles the Irritability facet from the lexical investigation by Saucier and Ostendorf.

The hierarchical representation of Emotional Stability and the intercorrelations among its three facets are presented in Figure 2.6 and Table 2.10, respectively. Note that unlike in the other Big Five domains, these intercorrelations all exceed .6, indicating that the Emotional Stability facets measure closely related phenomena. From selection and classification standpoints, these facets are therefore unlikely to have differential relations with organizational criteria. But, for developmental purposes, feedback beyond a single score for general Emotional Stability might be useful. For this reason, we chose a three-facet solution for TAPAS development, which includes *Even Tempered, Optimism, and Adjustment.*

Note that several scales such as the NEO Self-Consciousness and Vulnerability scales and the HPI Good Attachment scales appear factorially complex. This suggests that items constituting these scales tap into more than one Emotional Stability facet. This is not surprising because many of these scales were not developed to conform to the proposed three-facet representation.

Table 2.9. The 3-Facet EFA Solution for Emotional Stability

Scale Name	Emotional Stability Facets		
	Optimism	Adjustment	Even Tempered
HPI No Depression	**.81**	-.05	-.09
AB5C Moderation	**.79**	-.24	.21
NEO Depression	**-.75**	-.17	.07
HPI No Guilt	**.68**	.01	-.02
CPI Well-Being	**.62**	.09	.08
AB5C Happiness	**.62**	.30	.02
HPI Identity	**.61**	-.08	-.06
16PF Emotional Stability	**.61**	.13	.10
NEO Impulsiveness	**-.50**	.23	-.29
NEO Self-Consciousness	-.49	-.43	.21
NEO Vulnerability	-.48	-.34	.03
HPI Good Attachment	.45	-.27	.15
MPQ Alienation	-.39	.05	-.07
HPI No Somatic Complaints	.34	.17	-.03
JPI Cooperativeness	.27	**-.85**	.17
16PF Apprehensive	-.18	**-.70**	.15
JPI Anxiety	.06	**-.68**	-.28
NEO Anxiety	-.18	**-.66**	-.02
HPI Not Anxious	-.05	**.64**	.20
MPQ Stress Reaction	-.21	**-.56**	-.17
AB5C Toughness	.21	**.50**	.13
AB5C Cool-Headedness	-.29	.42	.27
AB5C Calmness	.04	-.08	**.92**
NEO Hostility	-.10	.08	**-.81**
HPI Empathy	-.19	.14	**.74**
HPI Even Tempered	.14	-.19	**.70**
AB5C Stability	.13	.21	**.63**
AB5C Tranquility	.07	.17	**.54**
HPI Calmness	.10	.28	.34
NEO Straightforwardness	-.39	.36	-.41

Note. N = 747.

Table 2.10. Correlations between the TAPAS Facets of Emotional Stability

Facet	Optimism	Adjustment	Even Tempered
Optimism	1.00		
Adjustment	.81	1.00	
Even Tempered	.72	.68	1.00

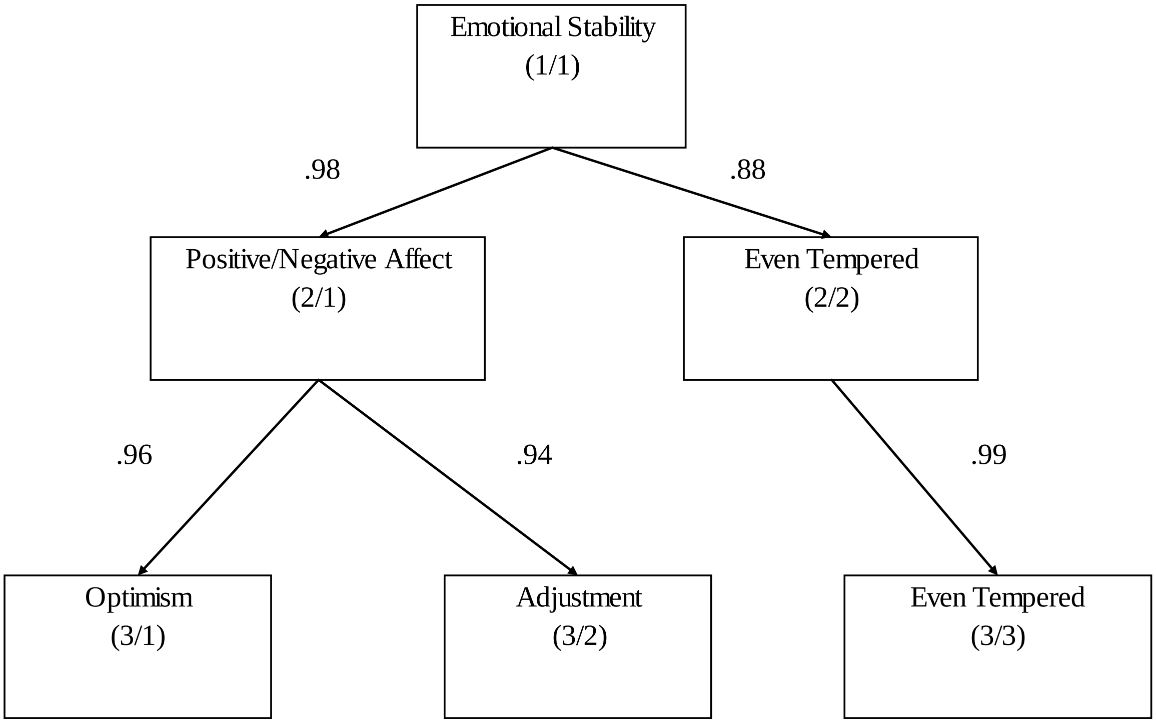

Figure 2.6. The Hierarchical Structure of Emotional Stability

Summary and Conclusions

The purpose of this chapter was to describe the development of the TAPAS trait taxonomy comprising a comprehensive set of nonredundant, narrow personality traits (facets) located within the Big Five framework. Depending on the purpose of assessment, TAPAS facet scores can serve as individual predictors of outcomes or be combined using statistical methods to form composites that are optimal for various applications. Rather than adhering to an existing rationale or theoretical nomenclature, our approach was to develop a narrow trait taxonomy for the Big Five by reviewing the factor analytic work of Saucier and Ostendorf (1999), who examined the factor structure of 500 adjectives describing human behavior (i.e., assertive, talkative, anxious). We then conducted our own factor analyses of a maximally diverse array of personality indicators drawn from seven widely used questionnaires. Our questionnaire-based factor analytic results corresponded closely with the structures derived by Saucier and Ostendorf.

A total of *22 personality facets* were identified and retained for TAPAS statement pool development (three to six facets per Big Five dimension). For each Big Five factor, we created a hierarchical structure showing the relationships among lower-order factors having various degrees of specificity. The last row in each hierarchical representation shows the facets we retained for TAPAS development purposes, while the correlations between factor scores at successive levels facilitates the mapping of connections between TAPAS facets and those in other existing personality inventories for construct and criterion-related validity investigations.

Table 2.11 summarizes the original TAPAS taxonomy. The table is organized into five broad clusters corresponding to the Big Five factors (see column 1). Within these clusters, each row shows a TAPAS facet name (column 2), examples of adjectives associated with the facet (column 3), and a brief description of a typical high scoring examinee.

Table 2.11. Facet Taxonomy for TAPAS: Trait Names, Markers, and Descriptions

Big Five Factor	TAPAS Facet	Key Adjectives	Brief Description
Extraversion	**Dominance**	assertive, direct, submissive, helpless	High scoring individuals are domineering, "take charge" and are often referred to by their peers as "natural leaders."
	Sociability	sociable, gregarious, talkative	High scoring individuals tend to seek out and initiate social interactions.
	Attention Seeking	loud, entertaining, dull, unexciting, shy	High scoring individuals tend to engage in behaviors that attract social attention. They are loud, loquacious, entertaining, and even boastful.
	Physical Conditioning	active, vigorous, fit, inactive, brisk	High scoring individuals tend to engage in activities to maintain their physical fitness and are more likely participate in vigorous sports or exercise.
Agreeableness	**Consideration**	compassionate, warm, cold, insensitive	High scoring individuals are affectionate, compassionate, sensitive, and caring.
	Selflessness	charitable, helpful, generous, stingy, selfish	High scoring individuals are generous with their time and resources.
	Cooperation	agreeable, cordial, trusting, uncooperative	High scoring individuals are pleasant, trusting, cordial, non-critical, and easy to get along with.
Conscientiousness	**Achievement**	ambitious, industrious, aimless	High scoring individuals are seen as hard working, ambitious, confident, and resourceful.
	Order	organized, neat, sloppy	High scoring individuals tend to organize tasks and activities and desire to maintain neat and clean surroundings.
	Self Control	controlled, deliberate, inconsistent	High scoring individuals tend to be cautious, levelheaded, able to delay gratification, and patient.
	Responsibility	prompt, irresponsible, unreliable	High scoring individuals are dependable, reliable, and make every effort to keep their promises.
	Non-Delinquency	rule-following, lawful, delinquent	High scoring individuals tend to comply with rules, customs, norms, and expectations, and they tend not to challenge authority.
	Virtue	honest, frank, misleading	High scoring individuals strive to adhere to standards of honesty, morality, and "good Samaritan" behavior.

Table 2.11. Facet Taxonomy for TAPAS: Trait Names, Markers, and Descriptions (cont'd)

Big Five Factor	TAPAS Facet	Key Adjectives	Brief Description
Emotional Stability	**Adjustment**	relaxed, certain, insecure, nervous	High scoring individuals are well adjusted, worry free, and handle stress well.
	Even Tempered	calm, composed, moody, hot-headed	High scoring individuals tend to be calm and stable. They don't often exhibit anger, hostility, or aggression.
	Optimism	happy, optimistic, depressed, dejected	High scoring individuals have a positive outlook on life and tend to experience joy and a sense of well-being.
Openness To Experience	**Intellectual Efficiency**	intelligent, analytical, knowledgeable,	High scoring individuals believe they process information and make decisions quickly; they see themselves (and they may be perceived by others) as knowledgeable, astute, or intellectual.
	Ingenuity	creative, inventive, unimaginative	High scoring individuals are inventive and can think "outside of the box."
	Curiosity	curious, perceptive, unobservant,	High scoring individuals are inquisitive and perceptive; they are interested in learning new information and attend courses and workshops whenever they can.
	Aesthetics	aesthetic, artistic, unsophisticated, unrefined	High scoring individuals appreciate various forms of art and music and participate in art-related activities more than most people.
	Tolerance	tolerant, broadminded, biased	High scoring individuals scoring are interested in other cultures and opinions that may differ from their own.
	Depth	introspective, reflective, shallow	High scoring individuals tend to examine their lives and exhibit behaviors associated with self- improvement.

CHAPTER 3: DEVELOPMENT AND CALIBRATION OF STATEMENT POOLS TO ASSESS TAPAS TRAITS

In this chapter, we describe the development of statement pools for each of the 22 TAPAS facets. The statements were administered to large samples of respondents under "honest" and "fake good" conditions. Data from the honest condition were used to estimate Generalized Graded Unfolding Model (GGUM; Roberts, Donoghue, & Laughlin, 2000) statement parameters; data from the fake good condition were used to compute social desirability parameters. GGUM statement parameters and social desirability parameters are needed to form the pairwise preference items appearing in TAPAS tests.

Development of TAPAS Statement Pools

TAPAS statement pool development followed a four-stage process. In Stage 1, content domains relevant to each facet were identified by examining the relevant psychological literature, as well as the content of available items representing scales found to have high loadings on that facet in the factor analysis investigation (see Chapter 2). For example, according to the research literature, people who score high on the Order facet of Conscientiousness tend to describe themselves as organized, meticulous, neat, and punctual. Scales measuring this facet are Orderliness and Perfectionism from the AB5C, the Order scale from the NEO-PI-R, the Perfectionism scale from the 16PF, and the Organization scale of the Jackson Personality Inventory – Revised. Examples of statements commonly found in these scales are "I like order" and "I leave my belongings lying around."

In Stage 2, we wrote 60-70 statements assessing behaviors, cognition, and affect for each TAPAS facet. These statements were written to span the respective trait continua, varying in extremity from low to high. Care was also taken to include statements reflecting moderation or neutrality because they help to distinguish between examinees having moderate trait levels and those having somewhat low or high trait levels. Not only does this practice broaden the variety of statements that can be presented to examinees, but also it helps to balance measurement precision all along the trait continuum, which is particularly helpful in a computerized adaptive testing environment.

To ensure that the statements for the TAPAS facets spanned their respective trait continua, Ph.D. faculty members in Industrial and Organizational Psychology served as subject matter experts (SMEs). They were asked to judge the location of each statement on a scale from 1 (low) to 7 (high). The average rating of the SMEs was used as a proxy for statement extremity to identify possible gaps in the distributions of statement pools before pretest data were collected and calibrated using the GGUM. For example, the Order statement, "I am incapable of planning ahead," was rated a 1.5 by the SMEs, indicating a very low level or Order, whereas the statement, "I keep detailed notes of important meetings and lectures," was rated a 6.5, indicating very high Order. Statements with extremity ratings that varied markedly across SMEs were subsequently rewritten or discarded.

In Stage 3, the statements were carefully edited. Grammar and punctuation were examined and corrected as needed. The reading level of the statements was also examined to ensure that they were accessible to individuals with a high school education.

In Stage 4, statements were reviewed for length, clarity, and sensitivity. Overly long statements were either edited to reduce length or discarded. Statements were again examined for clarity and some were modified to improve readability. Importantly, items were examined for sensitive content and some were removed following this review.

Estimating GGUM Parameters for TAPAS Statements

To estimate GGUM parameters needed for construction of MDPP items, TAPAS statements were administered to large representative samples of Army recruits. It is important to note that these new Soldiers should be more representative of the population to be assessed by TAPAS (i.e., applicants for enlistment) than experienced Soldiers. Pretesting began in November of 2005 and ended in April of 2008. Recruit volunteers were obtained at Fort Jackson, Fort Leonard Wood, and Fort Benning; all data collections complied with American Psychological Association ethical guidelines for research with human subjects. The breakdown of various samples and the number of statements pretested are shown in Table 3.1.

Table 3.1. Breakdown of Samples Used to Estimate GGUM Parameters for TAPAS Statements

Date	Number of Recruits	Pretest Site	Number of TAPAS Statements Pretested
November 2005	270	Fort Leonard Wood	225
February 2006	272	Fort Leonard Wood	150
March 2006	525	Fort Jackson	150
June 2006	588	Fort Jackson	225
August 2006	532	Fort Jackson	225
January 2007	456	Fort Jackson	221
January 2007	456	Fort Jackson	221
February 2007 Part 1	319	Fort Leonard Wood	221
February 2007 Part 2	385	Fort Leonard Wood	208
May 2007	429	Fort Jackson	200
June 2007	585	Fort Benning	210
February 2008	452	Fort Benning	320

Each data collection focused on pretesting statements representing multiple TAPAS facets (usually 6 to 10 at a time). Using fewer facets would have made the content of the questionnaires too repetitive, and thus increased the risk of unmotivated responding. Each questionnaire had multiple forms. In each form, there were about 15 to 30 statements per facet with five to seven statements appearing in multiple forms for IRT linking purposes. Recruits were asked to indicate their level of agreement with each personality statement using a 4-point response format, where 1 = strongly disagree, 2 = disagree, 3 = agree, and 4 = strongly agree. In "honest" testing conditions, respondents were instructed to respond as honestly and accurately as possible and they were reminded that their responses would never be reported to supervisors or recorded in personnel records. The directions for the Pretest Questionnaire and two sample items are shown in Appendix A.

The 4-point response format was used for data collection to facilitate tests of essential unidimensionality based on EFA. Polytomously scored statements violate the normality assumption of EFA to a lesser degree than dichotomously scored statements and, thus, paint a more accurate picture regarding the dimensionality of a statement set. Davison (1977) showed that responses consistent with a unidimensional ideal point (unfolding) model generally display two major principal components and that the component loadings will show a simplex pattern. In addition, Roberts et al. (2000) suggested that a statement can be considered unidimensional if its communality based on the first two principal components is greater than or equal to .3. These guidelines were followed to screen out TAPAS statements that did not adequately measure their intended facet. Because Stark’s multidimensional pairwise preference model requires GGUM parameters for dichotomously scored statements, we dichotomized the four-point polytomous responses after conducting unidimensionality checks. Specifically, “strongly disagree” and “disagree” responses were collapsed and recoded as 0s, while “strongly agree” and “agree” responses were recoded as 1s. The dichotomous case of the GGUM may be written as follows:

$$P[U_i = 1 | \theta_j] = \frac{\exp\left(\alpha_i\left[\left(\theta_j - \delta_i\right) - \tau_{i1}\right]\right) + \exp\left(\alpha_i\left[2\left(\theta_j - \delta_i\right) - \tau_{i1}\right]\right)}{1 + \exp\left(\alpha_i\left[3\left(\theta_j - \delta_i\right)\right]\right) + \exp\left(\alpha_i\left[\left(\theta_j - \delta_i\right) - \tau_{i1}\right]\right) + \exp\left(\alpha_i\left[2\left(\theta_j - \delta_i\right) - \tau_{i1}\right]\right)_i},$$

where $\theta_j =$ the location of respondent *j* on the continuum underlying responses, $\alpha_i =$ the discrimination parameter for statement *i*, $\delta_i =$ the location of statement *i* on the continuum underlying responses, and $\tau_{i1} =$ the location of the subjective response category threshold on the latent continuum. Statement parameters $(\alpha_i, \delta_i, \tau_{i1})$ were estimated using marginal maximum likelihood (MML; for a detailed description, see Bock & Atkin [1981]), which was implemented in the GGUM2000 and GGUM2004 computer programs by Roberts (2001) and Roberts, Fang, Cui, and Wang (2006).

Like many other IRT models, GGUM parameters for a given sample are estimated under the assumption that the distribution of person parameters (trait scores) is normal with a mean of zero and a standard deviation of one. Because statements from the same TAPAS facet were often calibrated using data collected from different samples of recruits, which could realistically differ somewhat in their trait distributions, statement parameters from different data collections had to be put on a common metric through a procedure known as linking. Essentially, mean location and mean discrimination parameters for statements appearing in common across forms were used to calculate linking constants and place the respective sets of statement parameters on a common scale. For more details about GGUM linking GGUM parameter estimates, see the GGUMLINK manual (Roberts, 2002).

Estimating Social Desirability Parameters for TAPAS Statements

As was discussed in Chapter 1 of this report, the basic idea of creating fake-resistant personality items entails pairing statements that are similar in social desirability. That implies that each statement has a social desirability parameter reflecting the likelihood that it will be endorsed or agreed with by respondents who are trying to fake good. As was noted by Chernyshenko, Stark, Prewett, Gray, Stilson, and Tuttle (2009), previous studies involving forced choice items explored a range of options for estimating social desirability. Some researchers asked judges to explicitly rate the desirability of statements (Jackson, Wrobleski, & Ashton, 2000), while others derived desirability ratings from self-report data collected under “fake good” instruction sets (White & Young, 1998).

Because there were no guidelines about which approach should be preferred, we decided to estimate the social desirability of TAPAS statements using both approaches. In February 2008, we asked 276 recruits, organized in groups of 30-40, to pretend they were recruiters, and their task was to indicate how impressed they would be if an applicant gave them “agree” responses to various statements; the rating scale was 1 = "Not at all impressed" to 5 = "Highly Impressed." In April 2008, 221 recruits, again organized in groups of 30-40, were given strong instructions to “fake good” on the same sets of TAPAS statements; the rating scale was 1 = “Strongly Disagree” to 4 = “Strongly Agree.” Next, we averaged the ratings for the respective statements across respondents to yield two sets of social desirability estimates for the TAPAS statements - one representing recruits acting as judges and another representing recruits acting as applicants. In total, desirability ratings for 1260 TAPAS statements were obtained.

The correlation between the two sets of social desirability ratings for the 1260 statements was .87. This indicated that the respondents acting as judges (recruiters) and the respondents acting as applicants saw desirability in a similar way; the ordering of statements in terms of desirability was essentially the same. The two approaches differed markedly, however, in terms of the administrative burden and examinee reactions. Having respondents fake good took considerably less time and elicited far fewer questions than the alternative judgment task. For these reasons, we retained the “fake good” social desirability ratings for MDPP test construction purposes and used that approach for all subsequent statement pool development efforts. The directions for the Social Desirability Questionnaire and two samples questions are shown in Appendix B.

Summary

In sum, over 1200 statements measuring 22 TAPAS facets were developed. These statements were written to reflect low, medium, and high locations on each trait continuum. They were pretested on large samples of Army recruits in three different U.S. Army installations over a period of three years. GGUM and social desirability parameters were estimated for each statement for the future construction of MDPP test forms. Statements having GGUM discrimination parameters below .50 were removed from the pool because they would be unlikely candidates for inclusion in MDPP tests. In total, this effort produced 985 usable statements for TAPAS; the detailed breakdown of the number of statements per TAPAS facet is shown in Table 3.2. Two example statements are also shown for each TAPAS facet - one statement with a positive location parameter and the other with a negative location parameter.

Concurrent with the TAPAS item pool development, ARI researchers wrote new statements to possibly augment the AIM inventory (see White & Young, 1998, for a description of the AIM). Several dozen statements were pretested at Fort Jackson and Fort Leonard Wood using the same samples of recruits used in TAPAS pool development. Because AIM statements could be straightforwardly mapped onto the TAPAS facets, a decision was made in 2008 to augment the TAPAS statement pool with the ARI statements. GGUM parameters for the ARI statements were estimated in the same manner as described above. And although social desirability parameters for the ARI statements were initially produced via the "judgment task," they were utilized for subsequent MDPP test construction purposes because of the .87 correlation between the judgment and fake good desirability ratings for the TAPAS statements. Altogether, 149 ARI statements measuring 9 facets were added to the form the initial statement pool for MDPP testing. A breakdown of the resulting statement pool for each TAPAS facet is presented in Table 3.2.

Table 3.2. Number of Statements Available for Each of the 22 TAPAS Facets

TAPAS Facet	TAPAS Pool	ARI Pool	Total Available	Examples of Statements with Positive and Negative Locations
Cooperation	45	17	62	*I am a really easy person to live with.* *I have often been critical of others.*
Selflessness	43		43	*I contribute to charity regularly.* *I only help people when I know I will get something in return.*
Consideration	48		48	*Most people would say that I am a loving and forgiving person.* *I can't stand listening to others complain about their problems, so people don't come to me for support.*
Achievement	53	22	75	*I try to be the best at anything I do.* *I finish tasks at my convenience.*
Non-Delinquency	34	17	51	*I support long-established rules and traditions.* *When I was in school, I used to break rules quite regularly.*
Order	41		41	*I am definitely more organized than most people.* *Others always tell me to clean up my work area.*
Responsibility	54		54	*I have made great personal sacrifices to do what I have promised.* *When things go wrong, I'd rather blame it on bad luck than admit that I may have been at fault.*
Self Control	56		56	*I am really good at tasks that require a careful and cautious approach.* *I often rush into action without thinking about the consequences.*
Virtue	40	8	48	*I firmly believe that under no circumstances is it okay to lie.* *I try to do the right thing, but sometimes it is necessary to cut some corners.*
Even Tempered	38	14	52	*Even during a particularly heated argument, I keep my emotions under control.* *People who know me well would say that I am moody.*
Adjustment	41	14	55	*Even if I've had a really stressful day at work, I fall asleep easily.* *Because I constantly worry about things, it is hard for me to relax.*

Table 3.2. Number of Statements Available for Each of the 22 TAPAS Facets (cont'd)

TAPAS Facet	TAPAS Pool	ARI Pool	Total Available	Examples of Statements with Positive and Negative Locations
Optimism	39	12	51	*I never get depressed.* *I have a hard time finding positive things to say about myself.*
Dominance	42	24	66	*After joining a group, I usually end up becoming the leader.* *I've been told that I need to be more assertive.*
Attention Seeking	49		49	*I like to be the center of attention.* *I don't like to be noticed.*
Physical Condition	64	21	85	*I like to exercise.* *I don't consider myself to be an athletic person.*
Sociability	40		40	*I'll talk to anyone.* *It takes a while to get to know me.*
Aesthetics	43		43	*I appreciate the paintings of well-known artists.* *I think viewing art is a waste of time.*
Curiosity	43		43	*I like to analyze things instead of taking them at face value.* *As long as I pass a test, I don't care what I have learned.*
Depth	50		50	*One of the main goals in life should be to understand its meaning.* *I try not to think too deeply about the future.*
Ingenuity	45		45	*Generating new ideas is effortless for me.* *I rarely take an idea and apply it in a new way.*
Intellectual Efficiency	40		40	*I am very quick at processing information.* *I usually struggle to solve complex problems.*
Tolerance	37		37	*I feel that an opportunity to learn about the culture of others is something to be treasured.* *I like visiting familiar places and avoid trips outside my country as best I can.*
Total Statements	**985**	**149**	**1134**	

CHAPTER 4: CONSTRUCT AND CRITERION-RELATED VALIDITIES OF TAPAS FACETS

This chapter presents several efforts conducted to investigate construct and criterion-related related validities of TAPAS facets. Because TAPAS was designed specifically for selection and classification purposes, we felt it was necessary to link the TAPAS facet taxonomy to criteria commonly used to evaluate the performance of military personnel. To do that, we first conducted a meta-analysis that looked at reported validities for broad and narrow personality traits for military, police, and fire-fighting personnel, and determined which TAPAS facets would be most relevant for predicting attrition, training performance, fitness levels, and other outcomes. Next, we collected data that allowed us to compare scores based on traditional personality scales (items administered in a single statement format) to those based on pairwise preference scales (items administered in unidimensional and multidimensional pairwise preference formats). Results showed that our IRT-based MDPP scores were highly comparable to traditional scores in terms of construct and criterion validities. Finally, we discuss a large scale project in which a paper-and-pencil MDPP test measuring 12 TAPAS facets was administered to new Army recruits undergoing their basic training. Results showed good construct validities for TAPAS facets when compared to other personality measures used by the military, as well as promising patterns of criterion validities, especially those not typically predicted well by cognitive ability tests (e.g., adjustment to military life, intentions to stay in the military, and physical fitness). This project is particularly important because it provides a necessary benchmark for comparing results from future operational TAPAS tests.

Meta-Analysis of TAPAS Facet Validities in Military Settings

The purpose of the meta-analysis was to provide a quantitative summary of existing investigations linking TAPAS traits to performance criteria relevant in military contexts. Based on facet structures presented in Chapter 2 of this report, we first mapped existing personality measures into the TAPAS nomenclature. Next, we identified 43 unique studies published between 1988 and 2008 that utilized personality scales to predict performance in military, police, or firefighter occupations and created a database of over 1500 criterion correlations for broad and narrow personality traits comprising TAPAS for eight criteria: task proficiency, contextual performance, counterproductivity, attrition, leadership, training performance, adaptability, and fitness level. Finally, we computed meta-analytic validity estimates by averaging the sample size-weighted observed correlations for each facet-criterion relation with and without corrections for unreliability and sampling error.

Identification of Studies

We searched for relevant articles and technical reports starting from the year 1988. In our view, that year represents the advent of the modern view of job performance as an explicitly multidimensional construct (Campbell, 1990). That was also the year when one of the most significant military investigations of Soldier performance, Project A, reported important findings. Potential articles for inclusion were identified by conducting both electronic and manual searches. First, a computer-based electronic search was conducted in ERIC, ProQuest, PsycARTICLES, PsycINFO, and ScienceDirect for the years 1988–2008, using a number of relevant keywords such as personality, five factor model, and the military. Unclassified published military technical reports were also obtained. A small number of articles and reports conducted with civilian occupations similar to military jobs (e.g., firefighters, police) also were sourced.

The obtained studies had to be judged as being of reasonable quality and could not use previously published data. In total, 43 data sources were found to satisfy our inclusion criteria, yielding 1608 correlation coefficients. The 43 studies included in the meta-analysis are listed in Table 4.1 below.

Table 4.1. Studies Used in the Meta-Analysis

1	Atwater, L. E., & Yammarino, F. J. (1993). Personal attributes as predictors of superiors' and subordinates' perceptions of military academy leadership. *Human Relations, 46, 5,* 645-668.
2	Atwater, L. E., Dionne, S. D., Avolio, B., Camobreco, J. F., & Lau, A. W. (1999). A longitudinal study of the leadership development process: Individual differences predicting leader effectiveness. *Human Relations, 52*, 1543-1562.
3	Barrick, M.R. & Mount, M.K. (1993). Autonomy as a Moderator of the Relationships Between the Big Five Personality Dimensions and Job Performance. *Journal of Applied Psychology, 78*, 111-118.
4	Bartone, P. T., Snook, S. A., & Tremble T. R. (2002). Cognitive and personality predictors of leader performance in West Point cadets. *Military Psychology, 14,* 321-338.
5	Bartram, D. (1995). The predictive validity of the EPI and the 16PF for military flying training. *Journal of Occupational and Organisational Psychology, 68*, 219-236.
6	Black, J. (2000). Personality testing and police selection: Utility of the Big Five. *New Zealand Journal of Psychology, 29,* 2-21.
7	Bradley, J. P., Nicol, A. A. M., Charbonneau, D., & Meyer, J. P. (2002). Personality correlates of leadership development in Canadian forces officer candidates. *Canadian Journal of Behavioural Sciences, 34*, 92-103.
8	Connelly, M. S., Gilbert J A., Zaccaro S J., Threlfall, K. V, Marks, M. A., & Mumford, M. D. (2000). Exploring the relationship of leadership skills and knowledge to leader performance. *Leadership Quarterly, 11,* 65-86.
9	Detrick P., Chibnall, J. T., & Luebbert (2005). Relationship between personality and academy performance. *Applied HRM Research, 10,* 99 - 102.
10	Dorner, K. R. (1991). *Personality characteristics and demographic variables as predictors of job performance in female traffic officers.* Unpublished doctoral dissertation. United States International University.

Table 4.1 Studies Used in the Meta-Analysis (cont'd)

11	Driskell J. E., Hogan J., Salas E., & Hoskin, B. (1994). Cognitive and personality predictors of training performance. *Military Psychology, 6,* 31-46.
12	Duffy, M.K., Ganster, D.C., & Shaw, J.D. (1998). Positive affectivity and negative outcomes: The role of tenure and job satisfaction. *Journal of Applied Psychology, 83,* 950-959.
13	Foti, R. J, & Hauenstein, N. M. A., (2007). Pattern and variable approaches in leadership emergence and effectiveness. *Journal of Applied Psychology, 92,* 347-355.
14	Halfhill, T., Nielsen, T., Sundstrom, E. & Weilbaecher, A. (2005). Group personality composition and performance in military service teams. *Military Psychology, 17,* 41-54.
15	Hartmann, E., Sunde, T., Kristensen, W., & Martinussen, M. (2003). Psychological measures as predictors of military training performance. *Journal of Personality Assessment, 80,* 87-98.
16	Hogan, J., & Hogan, R. (1989). Noncognitive predictors of performance during explosive ordinance disposal training. *Military Psychology, 1,* 117-133.
17	Hogan, J., Rybicki, S. L., Motowildo, S. J., & Borman, W. C. (1998). Relations between contextual performance, personality, and occupational advancement. *Human Performance, 11,* 189-207.
18	Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology Monograph, 75,* 581-595.
19	Houston, J. S., Borman, W. C., Farmer, W. L., & Bearden, R. M. (2005). *Development of the Enlisted Computer Adaptive Personality Scales (ENCAPS) for the United States Navy,* Phase 2 (Institute Report No. 503). Minneapolis, MN: Personnel Decisions Research Institutes.
20	Hwang, G. S. (1988). *Validity of the California Psychological Inventory for police selection.* Unpublished master's thesis: North Texas State University
21	Knapp, D. J., & Heffner, T. S. (Eds.). (2010). *Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
22	Knapp, D. J., McCloy, R. A., Heffner, T. S. (2004). *Validation of measures designed to maximize 21st-century Army NCO performance* (Technical Report 1145). HUMRRO: Alexandria, VA.
23	Lall, R., Holmes, E. K., Brinkmeyer, K. R., Johnson, W. B., & Yatko, B. R. (1999). Personality characteristics of future military leaders. *Military Medicine, 164,* 906-910.
24	Larson, G. E., Booth-Kewly S., & Ryan, M. (2002). Predictors of Navy attrition. II. A demonstration of potential usefulness for screening. *Military Medicine, 167,* 770-777.
25	Lyons, T. J., Bayless, A., & Park, R. K. (2001). Relationship of cognitive, biographical, and personality measures with the training and job performance of detention enforcement officers in a federal government agency. *Applied HRM Research, 6,* 67-70.
26	Mael, F. A., & Ashforth, B. E. (1995). Loyal from day one: Biodata, organizational identification, and turnover among newcomers. *Personnel Psychology, 48,* 309-333.

Table 4.1 Studies Used in the Meta-Analysis (cont'd)

27	Mael, F. A., & Hirsch, A. C. (1993). Rainforest empiricism and quasi-rationality: Two approaches to objective biodata. *Personnel Psychology, 46,*719-738.
28	McCormack, L., & Mellor, D. (2002). The role of personality in leadership: An application of the five-factor model in the Australian military. *Military Psychology, 14,* 179-197.
29	Milan, L. M. (2002). *Analog scales as temperament measures in the Baseline Officer Longitudinal data set (BOLDS).* (Technical Report 1126). Army Research Institute for Behavioral and Social Sciences: Alexandria, VA.
30	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.
31	Paunonen, S. V., Lönnqvist, J., Verkasalo, M., Leikas, S., & Nissinen, V. (2006). Narcissism and emergent leadership in military cadets. *Leadership Quarterly, 17,* 475-486.
32	Perkins, A. M., & Corr, P. J. (2006). Cognitive ability as a buffer to neuroticism: Churchill's secret weapon? *Personality and Individual Differences, 40,* 39-51. (Sample 1)
33	Ployhart, R. E., Lim, B. C., & Chan K. Y. (2001). Exploring relations between typical and maximal performance ratings and the five factor model of personality. *Personnel Psychology, 54,* 809 -843 .
34	Pugh, G. (1985). The California Psychological Inventory and police selection. *Journal of Police Science and Administration, 13,* 172-177.
35	Pulakos, E. D., Borman, W. C., & Hough, L. M. (1988). Test validation for scientific understanding: Two demonstrations of an approach to studying predictor-criterion linkages. *Personnel Psychology, 41,* 4, 703-716.
36	Siem, F. M. (1996). The use of response latencies to enhance self-report personality measures. *Military Psychology, 8,* 15-27.
37	Stricker, L. J., & Rock, D. A. (1998) Assessing leadership potential with a biographical measure of personality traits. *International Journal of Selection and Assessment, 6,* 164-184.
38	Strickland, W. J. (2005). *A longitudinal examination of first term attrition and reenlistment among FY1999 enlisted accessions* (Technical Report 1172). HUMRRO: Alexandria, VA.
39	Super, J. T. (1995). Psychological characteristics of successful SWAT/tactical response team personnel. *Journal of Police and Criminal Psychology, 11,* 60-63.
40	Surrette, M. A., Aamodt, M. G., & Serafino, G. (1990). *Validity of the New Mexico Police Selection Battery*. Paper presented at the annual meeting of the Society for Police and Criminal Psychology, Albuquerque, NM.
41	Van Scotter, J. R., & Motowildo, S. J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology, 81,* 525-531.
42	Vickers, R. R., Hervig. L. K., & Booth, R. F. (1996). *Personality and success among military enlisted personnel: An historical prospective study of US Navy corpsmen* (Report No. 96-15). San Diego: CA: Naval Health Research Center.
43	Wright, B. S., Doerner, W. G., & Speir, J. C. (1990). Pre-employment psychological testing as a predictor of police performance during an FTO program. *American Journal of Police, 9,* 65-83.

Criterion Type

The first three criteria, task performance, contextual performance, and counterproductivity, represent the three broad domains of performance suggested in the current industrial and organizational psychology literature (Dalal, 2005; Rotundo & Sackett, 2002; Sackett, 2002; Viswesvaran & Ones, 2000). Task performance is defined as “activities that contribute to the organization’s technical core either directly by implementing a part of its technological process, or indirectly by providing it with needed materials or services” (Borman & Motowidlo, 1997, p. 99). In this meta-analysis, task performance measures included supervisory ratings of general soldier proficiency, technical performance, and overall performance. On the other hand, contextual performance or organizational citizenship behavior (OCB) is usually defined as voluntary, discretionary behavior that is typically not recognized or rewarded but improves organizational effectiveness (Dalal, 2005; Organ, 1988). In military settings, measures of contextual performance include commendations, helping peers, working well with others, dedication, initiative, and work ethics ratings. Counterproductivity, or counterproductive work behavior (CWB), is considered to be undesirable, in that it represents employee behavior that is contrary to the organization’s legitimate interests and is intended to harm the organization (Dalal, 2005; Robinson & Bennett, 1995; Rotundo & Sackett, 2002). Examples include number of disciplinary incidents reported, ratings of personal discipline, integrity, and honesty.

The fourth criterion variable was turnover. In military settings, turnover takes the form of voluntary or involuntary attrition from the Service prior to contract completion. This is often seen as the most important criterion because it is very costly. Moreover, it is fairly easy to calculate its costs. Considering the high costs of training, even small reductions in attrition could result in significant monetary savings for the military (Stark, Chernyshenko, Drasgow, Lee, White, & Young, 2011).

The fifth criterion, leadership effectiveness, has been widely researched in relation to personality traits (e.g., Bono & Judge, 2004; Judge & Bono, 2000). Consistent with the current leadership literature, we define leadership effectiveness as leaders’ performance in guiding and motivating followers to achieve valued outcomes in the organization (Judge, Bono, Ilies, & Gerhardt, 2002). In our meta-analysis, we considered the following as measures of leadership performance: leader effectiveness/performance rated by subordinates, peers, and leaders themselves, military leadership grades of officer candidates, and transformational leadership scores as measured by the Multifactor Leadership Questionnaire (Avolio & Bass, 2002), and other similar measures of effective leadership behaviors.

The last three criteria, training performance, adaptability, and fitness level, are somewhat narrow in scope and are particularly relevant in the military. Results from job knowledge tests, training grades, adaptability, and adjustment ratings by peers, supervisors, and oneself, as well as the most recent Army Physical Fitness Test (APFT) scores or ratings of fitness levels were used as measures of these important criteria.

Analyses

A meta-analysis procedure recommended by Hunter and Schmidt (2004) was used to quantitatively summarize previous research findings between personality measures and the eight aforementioned organizational criteria. We computed the average correlation across individual studies weighted by sample size and corrected for measurement errors (i.e., unreliability) in both the predictor and criterion variables. The unreliabilities were corrected by using information from individual studies when it was available. When studies did not report local reliabilities for the predictor and/or criterion measures, we used a conservative value of .80 for both so that we would not overestimate the magnitude of the true predictor-criterion correlation by overcorrection.

Results

Tables 4.2-4.9 present meta-analytic results for the 22 TAPAS facets and 8 criteria: task performance (Table 4.2), contextual performance (Table 4.3), counterproductivity (Table 4.4), training performance (Table 4.5), leadership effectiveness (Table 4.6), turnover (Table 4.7), adaptability (Table 4.8), and fitness performance (Table 4.9). We also show results at the broad factor level (a.k.a., Big Five), in which correlations for facets representing a particular broad factor were aggregated. Note that some studies used scales deemed to measure only broad factors, so the number of correlations for each broad factor was higher than the total number of correlations across facets.

Results in each table are organized as follows. Column 1 shows the 22 TAPAS facets grouped beneath the appropriate the Big Five factor. Columns 2, 3, and 4 show the total sample size, the number of studies, and the number of criterion correlations that were aggregated. Columns 5, 6, and 7 show the average observed correlations weighted by sample size and the associated 95 % confidence intervals. The last two columns show the reliability-corrected validities and their estimated standard deviations.

As can be seen in Table 4.2, validities of various personality dimensions for predicting task performance were not particularly high. Most uncorrected validities were in the .05 to .15 range, with facets of Emotional Stability and Conscientiousness having higher validities than facets of Openness and Extraversion. The magnitudes of these validities were in line with previous research showing that personality constructs generally have lower validities for predicting task performance than cognitive ability constructs.

Meta-analytic validities for predicting contextual performance, shown in Table 4.3, were higher than those for task performance. Several uncorrected validities reached or exceeded .20, which is considered high for personality predictors. The best predictors of contextual performance were the Order and Achievement facets of Conscientiousness and the Adjustment and Even Tempered facets of Emotional Stability. Extraversion facets also showed noteworthy relationships with contextual performance. Dominance and Physical Conditioning (Activity) facets had positive correlations, while Attention Seeking had a negative correlation with contextual performance.

The best predictors of counterproductivity were facets of Conscientiousness (Non-Delinquency, Achievement, Self Control, and Order) and Agreeableness (Cooperation). All had negative relations with counterproductivity (see Table 4.4). As expected, Attention Seeking, a facet of Extraversion, had a positive relationship with counterproductivity, meaning that high scoring individuals tended to have more disciplinary problems. Somewhat surprisingly, Emotional Stability facets had negligible correlations with counterproductivity. Only one of its facets, Even Tempered, had an uncorrected validity reaching -.10.

Training performance was best predicted by Intellectual Efficiency, Curiosity, Achievement, Optimism, and Dominance (see Table 4.5). Together these dimensions describe goal oriented individuals with a positive outlook on life and an interest in intellectual endeavors. Although the validities are not very high, the results indicate that training performance can be enhanced by selecting individuals who are not just capable of learning but who are also motivated to learn.

Facet level validity data for Leadership Effectiveness was rather sparse. Most studies focused on Big Five measures. Extraversion had the highest uncorrected validity (.16), followed by Conscientiousness (.13), Openness (.13) and Emotional Stability (.12). At the facet level, Responsibility was the best personality predictor, with an uncorrected validity estimate of .24. Overall, similar to what has been shown in civilian meta-analyses (see Judge & Bono, 2000), personality dimensions appear to play an important role in predicting leadership performance in military settings.

Our results also show that personality predicts turnover, which is a very important criterion for organizations faced with high training costs. Facets of Emotional Stability and Conscientiousness were generally negatively associated with this turnover (Table 4.7). And, although perhaps specific to the military, the Physical Conditioning (Activity) facet predicted turnover: The uncorrected validity estimate was -.14. This is not very surprising considering the high physical demands often placed on enlisted personnel and officers. Curiosity was also negatively correlated with turnover, which was again expected given the continuous training demands of military jobs.

Adaptability was best predicted by Emotional Stability, Extraversion, and Conscientiousness facets. The highest uncorrected validity was for Adjustment, followed by Achievement and Physical Conditioning (Activity). Soldiers' fitness levels were also predicted by Extraversion and Conscientiousness facets. In fact, Physical Conditioning (Activity) scores had the highest uncorrected validity of .27.

In sum, the results of our meta-analysis identified several TAPAS facets that predicted one or more criteria that are important to the military. These included the Dominance, Physical Conditioning, and Attention Seeking facets of Extraversion, the Cooperation facet of Agreeableness, the Optimism, Adjustment and Even Tempered facets of Emotional Stability, the Achievement, Order, Non-Delinquency, and Responsibility facets of Conscientiousness, and the Intellectual Efficiency and Curiosity facets of Openness. These facets are recommended for use by the military to enhance the quality of selection and classification decisions.

Table 4.2. Meta-Analytic Results for TAPAS Facets and Task Performance

Personality Dimension	**N**	**k_d**	**k_c**	**r_{xy}**	**95 % CI**		**$\rho_{xy(b)}$**	**Sres(b)**
					lower	**upper**		
Extraversion	60382	29	71	.02	-.05	.08	.02	.06
Sociability	4213	8	31	.04	-.13	.21	.05	.07
Dominance	18944	11	17	.02	-.04	.08	.02	.04
Attention Seeking	1607	3	5	.05	-.06	.16	.07	.12
Physical Cond. (Activity)	33174	4	10	.00	-.03	.04	.00	.06
Agreeableness	25289	17	39	.04	-.04	.11	.04	.04
Cooperation	19029	5	14	.02	-.03	.08	.03	.02
Consideration	1971	2	7	.07	-.04	.19	.09	.00
Selflessness	499	5	6	.14	-.07	.35	.17	.00
Emotional Stability	24539	21	38	.05	-.03	.13	.06	.06
Optimism	1937	8	11	.10	-.05	.25	.12	.07
Adjustment	3582	4	12	.10	-.02	.21	.12	.09
Even Tempered	1933	4	6	.09	-.02	.20	.11	.10
Conscientiousness	96183	42	100	.05	-.01	.12	.07	.07
Achievement	37519	12	30	.04	-.01	.10	.05	.07
Order	2404	4	9	.16	.04	.28	.20	.11
Responsibility	2031	6	20	.09	-.10	.28	.12	.00
Non-Delinquency	32965	4	10	.06	.02	.09	.07	.03
Self Control	1323	3	6	.13	.00	.26	.17	.07
Virtue	1646	7	14	.07	-.11	.25	.09	.00
Openness	12141	18	56	.06	-.07	.19	.07	.03
Intellectual Efficiency	2653	7	11	.08	-.04	.21	.10	.07
Ingenuity	-	-	-	-	-	-	-	-
Curiosity	895	2	3	.06	-.06	.17	.07	.04
Aesthetics	1704	1	6	.05	-.06	.17	.06	.00
Tolerance	5745	8	31	.04	-.10	.19	.05	.03
Depth	-	-	-	-	-	-	-	-

Table 4.3. Meta-Analytic Results for TAPAS Facets and Contextual Performance

Personality Dimension	**N**	**k_d**	**k_c**	**r_{xy}**	**95 % CI**		**$\rho_{xy(b)}$**	**Sres(b)**
					lower	**upper**		
Extraversion	34104	14	44	.12	.05	.19	.15	.11
Sociability	2494	4	17	-.01	-.17	.15	-.01	.07
Dominance	12392	6	15	.13	.07	.20	.17	.05
Attention Seeking	1068	1	4	-.13	-.25	-.01	-.18	.22
Physical Cond. (Activity)	17134	2	6	.14	.11	.18	.18	.09
Agreeableness	15304	9	22	.11	.03	.18	.14	.08
Cooperation	9723	3	6	.13	.08	.18	.17	.06
Consideration	534	1	2	.09	-.03	.21	.14	.08
Selflessness	333	2	3	-.05	-.24	.13	-.07	.13
Emotional Stability	11320	7	16	.14	.07	.21	.18	.07
Optimism	333	2	3	.08	-.10	.27	.10	.00
Adjustment	585	1	3	.17	.03	.31	.21	.03
Even Tempered	585	1	3	.20	.06	.33	.24	.00
Conscientiousness	59473	24	88	.15	.08	.23	.20	.09
Achievement	19423	7	18	.21	.15	.26	.26	.06
Order	2874	2	14	.20	.07	.34	.26	.10
Responsibility	2067	3	13	.12	-.03	.27	.17	.17
Non-Delinquency	22161	4	21	.11	.05	.17	.14	.04
Self Control	1170	1	6	.14	.00	.28	.18	.00
Virtue	3002	4	12	.06	-.06	.18	.08	.14
Openness	5662	8	30	.06	-.08	.20	.08	.08
Intellectual Efficiency	427	3	4	-.07	-.26	.11	-.09	.07
Ingenuity	-	-	-	-	-	-	-	-
Curiosity	1170	1	6	.10	-.04	.24	.12	.00
Aesthetics	-	-	-	-	-	-	-	-
Tolerance	2169	3	15	.11	-.06	.27	.13	.06
Depth	-	-	-	-	-	-	-	-

Table 4.4. Meta-Analytic Results for TAPAS Facets and Counterproductivity

Personality Dimension	**N**	**k_d**	**k_c**	**r_{xy}**	**95 % CI**		***$\rho_{xy(b)}$***	**Sres(b)**
					lower	**upper**		
Extraversion	31737	17	32	-.03	-.09	.03	-.04	.09
Sociability	1344	3	6	.05	-.08	.18	.06	.00
Dominance	11372	5	11	-.02	-.08	.04	-.02	.04
Attention Seeking	653	2	3	.13	.00	.26	.15	.00
Physical Cond. (Activity)	18039	5	9	-.05	-.10	-.01	-.07	.10
Agreeableness	12519	10	25	-.15	-.24	-.06	-.19	.10
Cooperation	9516	4	10	-.18	-.24	-.12	-.23	.08
Consideration	296	1	4	-.04	-.27	.19	-.05	.00
Selflessness	98	1	1	-.01	-.21	.19	-.01	.00
Emotional Stability	12409	14	25	-.08	-.17	.00	-.11	.07
Optimism	899	3	6	.01	-.15	.17	.02	.00
Adjustment	1140	3	4	-.04	-.15	.08	-.05	.07
Even Tempered	848	3	4	-.10	-.23	.04	-.11	.06
Conscientiousness	50385	21	53	-.18	-.24	-.12	-.23	.10
Achievement	18971	6	18	-.13	-.19	-.07	-.17	.08
Order	1433	3	7	-.12	-.26	.01	-.15	.04
Responsibility	294	1	3	-.01	-.21	.19	-.02	.00
Non-Delinquency	20083	4	13	-.23	-.28	-.18	-.29	.10
Self Control	538	2	4	-.12	-.28	.05	-.15	.00
Virtue	1097	2	4	-.09	-.21	.03	-.12	.00
Openness	6528	12	31	-.01	-.15	.12	-.02	.00
Intellectual Efficiency	1555	3	4	.04	-.06	.14	.04	.00
Ingenuity	-	-	-	-	-	-	-	-
Curiosity	895	2	3	-.06	-.17	.05	-.07	.00
Aesthetics	296	1	4	-.03	-.26	.20	-.03	.00
Tolerance	2282	4	11	-.03	-.16	.11	-.03	.00
Depth	-	-	-	-	-	-	-	-

Table 4.5. Meta-Analytic Results for TAPAS Facets and Training Performance

Personality Dimension	**N**	**k_d**	**k_c**	**r_{xy}**	**95 % CI**		**$\rho_{xy(b)}$**	**Sres(b)**
					Lower	**Upper**		
Extraversion	31879	32	95	.06	-.05	.16	.07	.08
Sociability	6050	7	28	.03	-.10	.16	.04	.04
Dominance	11301	10	30	.11	.01	.21	.14	.06
Attention Seeking	4256	4	10	-.02	-.11	.08	-.02	.00
Physical Conditioning (Activity)	7498	6	18	.04	-.05	.14	.05	.05
Agreeableness	18559	19	61	.02	-.09	.14	.03	.07
Cooperation	7628	7	25	.04	-.08	.15	.05	.07
Consideration	2432	3	9	-.02	-.14	.09	-.03	.06
Selflessness	676	2	5	-.08	-.25	.09	-.10	.06
Emotional Stability	21039	24	63	.08	-.03	.18	.10	.07
Optimism	6023	7	20	.11	.00	.22	.14	.06
Adjustment	4804	4	13	.06	-.04	.16	.08	.06
Even Tempered	3858	5	10	.05	-.05	.15	.06	.10
Conscientiousness	38844	35	126	.08	-.04	.19	.10	.11
Achievement	12575	10	42	.11	.00	.22	.14	.13
Order	1936	3	6	.06	-.05	.17	.07	.06
Responsibility	2608	3	16	.05	-.10	.20	.06	.07
Non-Delinquency	10872	4	26	.07	-.03	.16	.09	.07
Self Control	2086	3	6	.08	-.02	.19	.10	.03
Virtue	5076	5	18	-.01	-.13	.10	-.01	.06
Openness	23728	22	75	.08	-.03	.19	.10	.13
Intellectual Efficiency	6027	6	16	.14	.04	.24	.17	.15
Ingenuity	580	1	1	.08	.00	.16	.10	.00
Curiosity	3330	2	6	.13	.05	.21	.16	.10
Aesthetics	2432	3	9	.01	-.11	.13	.02	.02
Tolerance	7336	5	30	.07	-.05	.20	.09	.12
Depth	-	-	-	-	-	-	-	-

Table 4.6. Meta-Analytic Results for TAPAS Facets and Leadership Effectiveness

Personality Dimension	N	k_d	k_c	r_{xy}	95 % CI		$\rho_{xy(b)}$	Sres(b)
					lower	upper		
Extraversion	24883	14	44	.16	.08	.24	.20	.11
Sociability	428	1	4	-.07	-.26	.12	-.08	.00
Dominance	8360	5	17	.11	.03	.20	.14	.08
Attention Seeking	-	-	-	-	-	-	-	-
Physical Cond. (Activity)	5995	3	11	.13	.05	.21	.16	.05
Agreeableness	14126	9	25	.08	.00	.16	.11	.07
Cooperation	1597	2	6	.04	-.08	.16	.05	.04
Consideration	428	1	4	.07	-.12	.26	.09	.00
Selflessness	-	-	-	-	-	-	-	-
Emotional Stability	16519	10	24	.12	.05	.20	.15	.06
Optimism	81	1	1	-.04	-.26	.18	-.05	.00
Adjustment	-	-	-	-	-	-	-	-
Even Tempered	-	-	-	-	-	-	-	-
Conscientiousness	31168	20	60	.13	.04	.21	.16	.08
Achievement	9429	7	20	.15	.06	.24	.18	.11
Order	-	-	-	-	-	-	-	-
Responsibility	1383	1	3	.24	.15	.32	.30	.00
Non-Delinquency	8762	4	20	.09	-.01	.18	.12	.08
Self Control	81	1	1	-.06	-.28	.16	-.07	.00
Virtue	901	1	2	.07	-.02	.16	.09	.00
Openness	12667	7	16	.13	.06	.20	.17	.10
Intellectual Efficiency	-	-	-	-	-	-	-	-
Ingenuity	-	-	-	-	-	-	-	-
Curiosity	-	-	-	-	-	-	-	-
Aesthetics	-	-	-	-	-	-	-	-
Tolerance	-	-	-	-	-	-	-	-
Depth	-	-	-	-	-	-	-	-

Table 4.7. Meta-Analytic Results for TAPAS Facets and Turnover

Personality Dimension	N	k_d	k_c	r_{xy}	95 % CI		$\rho_{xy(b)}$	Sres(b)
					Lower	Upper		
Extraversion	52383	25	40	-.07	-.13	-.02	-.08	.09
Sociability	1015	2	3	-.08	-.18	.03	-.09	.00
Dominance	19747	6	13	.00	-.05	.05	.00	.01
Attention Seeking	978	4	7	.00	-.16	.17	.01	.15
Physical Conditioning (Activity)	24881	6	10	-.14	-.18	-.10	-.16	.06
Agreeableness	28728	13	21	-.09	-.14	-.03	-.10	.03
Cooperation	5716	5	8	-.07	-.15	.00	-.08	.03
Consideration	366	2	4	-.03	-.24	.17	-.04	.00
Selflessness	4512	1	1	-.07	-.10	-.04	-.08	.00
Emotional Stability	93256	19	30	-.19	-.22	-.15	-.21	.06
Optimism	67742	5	9	-.22	-.24	-.20	-.24	.01
Adjustment	1234	4	6	-.09	-.22	.05	-.10	.00
Even Tempered	579	2	2	-.11	-.22	.01	-.12	.00
Conscientiousness	116505	18	35	-.15	-.18	-.12	-.17	.06
Achievement	19570	4	11	-.09	-.13	-.04	-.10	.00
Order	5236	4	5	-.08	-.14	-.02	-.09	.07
Responsibility	4512	1	1	-.07	-.10	-.04	-.08	.00
Non-Delinquency	19492	4	9	-.12	-.16	-.08	-.13	.05
Self Control	364	2	5	.17	-.06	.39	.19	.17
Virtue	66690	1	1	-.19	-.20	-.18	-.21	.00
Openness	5977	14	24	-.07	-.19	.06	-.08	.05
Intellectual Efficiency	1302	1	2	-.05	-.13	.03	-.06	.00
Ingenuity	-	-	-	-	-	-	-	-
Curiosity	1925	4	7	-.14	-.26	-.02	-.16	.04
Aesthetics	293	2	4	.05	-.18	.28	.06	.00
Tolerance	1559	3	5	-.04	-.15	.07	-.05	.00
Depth	109	1	1	.02	-.17	.21	.02	.00

Table 4.8. Meta-Analytic Results for TAPAS Facets and Adaptability

Personality Dimension	N	k_d	k_c	r_{xy}	95 % CI		$\rho_{xy(b)}$	Sres(b)
					lower	upper		
Extraversion	7321	4	14	.14	.06	.23	.17	.08
Sociability	797	1	1	.15	.08	.22	.18	.00
Dominance	3460	2	7	.15	.06	.23	.18	.05
Attention Seeking	505	1	1	.01	-.08	.10	.01	.00
Physical Cond. (Activity)	2559	2	5	.17	.08	.25	.20	.11
Agreeableness	3564	4	8	.05	-.05	.14	.06	.04
Cooperation	1406	2	3	.03	-.06	.12	.04	.08
Consideration	-	-	-	-	-	-	-	-
Selflessness	-	-	-	-	-	-	-	-
Emotional Stability	3454	6	8	.16	.06	.25	.19	.07
Optimism	505	1	1	.14	.05	.23	.16	.00
Adjustment	992	2	2	.23	.15	.32	.29	.00
Even Tempered	700	2	2	.12	.02	.23	.14	.06
Conscientiousness	10276	10	27	.11	.01	.21	.14	.08
Achievement	2949	3	7	.19	.10	.28	.23	.05
Order	1285	2	5	.14	.02	.26	.17	.08
Responsibility	-	-	-	-	-	-	-	-
Non-Delinquency	4751	3	11	.06	-.03	.16	.08	.05
Self Control	390	1	2	.13	-.01	.26	.16	.00
Virtue	901	1	2	.07	-.03	.16	.08	.00
Openness	4790	5	11	.09	-.01	.18	.11	.05
Intellectual Efficiency	1302	1	2	.10	.03	.18	.12	.04
Ingenuity	-	-	-	-	-	-	-	-
Curiosity	895	2	3	.06	-.05	.17	.07	.00
Aesthetics	-	-	-	-	-	-	-	-
Tolerance	1692	2	4	.14	.05	.23	.17	.00
Depth	-	-	-	-	-	-	-	-

Table 4.9. Meta-Analytic Results for TAPAS Facets and Fitness Performance

Personality Dimension	N	k_d	k_c	r_{xy}	95 % CI		$\rho_{xy(b)}$	Sres(b)
					lower	upper		
Extraversion	30482	19	28	.21	.16	.27	.27	.10
Sociability	1081	2	2	.07	-.01	.15	.08	.00
Dominance	10083	5	9	.16	.10	.21	.20	.06
Attention Seeking	863	3	3	.06	-.06	.17	.07	.10
Physical Cond. (Activity)	17749	6	10	.27	.22	.31	.34	.04
Agreeableness	11729	10	21	.08	.00	.16	.11	.10
Cooperation	9941	5	12	.10	.03	.17	.13	.09
Consideration	716	2	4	-.05	-.19	.10	-.06	.00
Selflessness	-	-	-	-	-	-	-	-
Emotional Stability	12452	14	21	.10	.02	.18	.13	.10
Optimism	937	3	4	-.04	-.17	.09	-.05	.00
Adjustment	1723	3	5	.02	-.08	.13	.03	.04
Even Tempered	863	3	3	-.03	-.15	.08	-.04	.00
Conscientiousness	47565	19	42	.16	.10	.22	.20	.08
Achievement	19702	6	20	.18	.12	.24	.22	.07
Order	863	3	3	.06	-.06	.17	.07	.00
Responsibility	-	-	-	-	-	-	-	-
Non-Delinquency	18034	3	11	.13	.08	.18	.17	.06
Self Control	358	2	2	-.08	-.23	.06	-.10	.09
Virtue	-	-	-	-	-	-	-	-
Openness	5615	9	19	-.03	-.14	.08	-.04	.07
Intellectual Efficiency	1586	2	3	.04	-.05	.12	.05	.00
Ingenuity	-	-	-	-	-	-	-	-
Curiosity	505	1	1	.03	-.06	.12	.03	.00
Aesthetics	716	2	4	-.09	-.24	.05	-.11	.00
Tolerance	2302	3	7	-.05	-.16	.06	-.06	.05
Depth	-	-	-	-	-	-	-	-

Response Format Research

The basic question that our response format research sought to answer was to what extent scores obtained from real people via pairwise preference formats would correlate with scores from the traditional single statement format. The simulations described in Chapter 1 indicated the viability of our proposed IRT methodology, but we needed to determine whether the MDPP scores would provide meaningful relationships with other variables when administered to research participants under controlled conditions, as a first step toward field testing. Theoretically, in settings where individuals are not motivated to distort their responses, scores from different formats collected at the same time should correspond highly.

To address the score comparability question, we collected data from 602 university participants who were asked to complete a questionnaire involving personality items presented in three formats: 1) single statement (SS) Likert-type items, 2) unidimensional pairwise preference (UPP) items, and 3) multidimensional pairwise preference (MDPP) items. The items assessed three personality facets: Order, Self-Control, and Sociability. In addition, participants were asked to complete health (a shortened version of the Health Behavior Checklist; Vickers, Conway, & Hervig, 1990) and behavioral (the Study Behavior Questionnaire; B. W. Roberts, 2002) checklists so that we could compare criterion-related validities. The instructions explicitly stated that the goal of the study was to explore a new format for assessing personality and emphasized the importance of answering all items honestly.

The first section of the questionnaire consisted of 36 SS personality items (12 per facet) that were administered using a four-point format (Strongly Disagree, Disagree, Agree, and Strongly Agree). The second section of the questionnaire consisted of 36 UPP personality items. Subject matter expert (SME) ratings of statement location were used to construct and score the UPP items, in the same manner that UPP tests were created and scored for a study involving the Computerized Adaptive Rating Scales (CARS) asssessment (Borman, Buck, Hanson, Motowidlo, Stark, & Drasgow, 2001) and the NCAPS assessment (Houston et al., 2005). Items for the MDPP section were developed as follows. First, SMEs rated the social desirability of each personality statement individually on a scale of 1 to 7. All SMEs had masters or doctorate degrees in industrial and organizational psychology and had extensive experiences in item writing and scale development processes. Next, statements similar in desirability (e.g., differing by less than 1.0), but representing different facets, were paired to form 30 multidimensional items. We then created 6 unidimensional items (2 per facet) to fix the scale (these items were different from those appearing in the UPP section). When completing the UPP and MDPP sections of the questionnaire, respondents were instructed to select the statement in each pair that was “More like me."

Responses to the SS measures were analyzed using the GGUM2000 computer program (Roberts, Donoghue, & Lauglin, 2000), which computes statement parameters via marginal maximum likelihood (MML) estimation and person parameters by expected a posteriori (EAP) estimation. The UPP measures were scored using Stark’s (2006) ZG-EAP program for the Zinnes and Griggs (1974) IRT model. Finally, the MDPP measure was scored using Stark’s (2002) program for dichotomous multi-unidimensional pairwise preference responses. Marginal reliabilities, trait score intercorrelations, and validities for predicting behavioral criteria were also computed. These findings are summarized in Table 4.10.

Table 4.10. Correlations between Personality Facet Scores obtained using Single Statement, Unidimensional Pairwise Preference, and Multidimensional Pairwise Preference Formats

Format	Facet	Response Format								
		SS			UPP			MDPP		
		Order	Self Control	Sociability	Order	Self Control	Sociability	Order	Self Control	Sociability
SS	Order	**.74**								
	Self Control	.26	**.54**							
	Sociability	-.09	-.23	**.62**						
UPP	Order	**.75**	.30	-.12	**.75**					
	Self Control	.24	**.55**	-.27	.28	**.53**				
	Sociability	-.06	-.28	**.76**	-.12	-.30	**.78**			
MDPP	Order	**.75**	.32	-.20	**.74**	.31	-.18	**.75**		
	Self Control	.19	**.54**	-.27	.25	**.62**	-.31	.34	**.66**	
	Sociability	-.10	-.13	**.75**	-.13	-.18	**.73**	-.14	-.17	**.73**

Note. N = 602; SS = single statement; UPP = unidimensional pairwise preference; MDPP = multidimensional pairwise preference. Reliability estimates appear in bold on the main diagonal.

Table 4.10 presents the correlations among Order, Self Control, and Sociability scores obtained using the SS, UPP, and MDPP formats. The values appearing on the main diagaonal are marginal reliabilities. The monotrait-heteromethod (Campbell & Fiske, 1959) correlations (shown in bold print) are nearly identical to the respective marginal reliabilities, indicating good convergent validity across formats. Moreover, similar intercorrelations and criterion-related validities were observed across formats, indicating that the MDPP, UPP, and SS measures yielded higly comparable scores. The observed correlation between Substance Avoidance and Sociability, for example, was -.18 across all three formats. Importantly, the correlations between Order and Self Control (both facets of Conscientiousness) were positive and similar in magnitude (about .38) across formats. Thus, in contrast to historical findings of negative intercorrelations among scores derived from MDPP measures due to ipsativity (Meade, 2004), both IRT methods for constructing and scoring pairwise preference tests yielded results that were essentially equivalent to those obtained with SS personality tests. Results of this laboratory investigation provided clear empirical support for the IRT-based pairwise preference test construction and scoring approaches (for more information, see Chernyshenko, Stark, Prewett, Gray, Stilson, & Tuttle, 2009).

The previous simulation studies showed accurate recovery of trait scores under a wide range of conditions, but questions remained as to whether violations of model assumptions would adversely affect the accuracy or relational equivalence of trait score estimates for real people. This investigation showed unequivocally that they did not. In fact, the findings support the use of both MDPP and UPP methods as alternatives to traditional single statement personality measures when respondents can reasonably be expected to answer honestly.

Field Research

The third source of construct and criterion validity evidence for TAPAS comes from the Expanded Enlistment Eligibility Metrics (EEEM) research (Knapp & Heffner, 2010). For that investigation, we constructed a MDPP paper-and-pencil personality test called TAPAS-95s (*s* stands for static or nonadaptive) measuring 12 facets of personality with 95 pairwise preference items. TAPAS-95s was administered along with several other noncognitive instruments in an effort to evaluate their potential use for selection and classification. TAPAS-95s was developed using statements from the TAPAS pool for which social desirability ratings and GGUM parameters were estimated, as was described in Chapter 3 of this report. Items constructed 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 an updated version of Stark's (2002) computer program for MDPP trait score estimation.

Several thousand Soldiers from six military occupational specialties (MOS) were followed through basic training and several criterion measures were collected, including scores on job-specific knowledge tests, self-reported scores on the Army Physical Fitness Test, and ratings of job satisfaction and career intentions. Soldiers were also evaluated by their peers and supervisors on several performance rating scales. The usefulness of TAPAS scores for predicting these criteria was then evaluated in comparison with other cognitive and noncognitive predictors developed by the Army.

EEEM Construct Validity Results

Table 4.11 shows correlations between TAPAS-95s facets and those assessed by two other personality inventories: the Assessment of Individual Motivation (AIM; White & Young, 1998) and the Rational Biographical Inventory (RBI; Kilcullen, Putka, McCloy, & Van Iddekinge, 2005). The AIM, which measures six broad personality dimensions predictive of first-term Soldier attrition and performance, uses a forced-choice tetrad format (i.e., each item consists of four statements). The RBI measures multiple personality or motivational characteristics important to entry-level Soldier performance and retention (Kilcullen et al., 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 they have engaged in physical activities) using multiple Likert-type response scales. Also shown are correlations between TAPAS-95s facets and the AFQT, which is a composite of the Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Math Knowledge subtests of the ASVAB.

As can be seen in Table 4.11, TAPAS-95s facets showed good construct validity. Intellectual Efficiency and Curiosity, for example, showed correlations of .38 and .24, respectively, with AFQT. This was expected, given that both facets tap the intellectance aspects of Openness to Experience, which is known to correlate with cognitive ability. The Intellectual

Efficiency and Curiosity facets also correlated with the RBI Cognitive Flexibility scale (.33 and .41, respectively), which was also designed as a measure of Openness. The TAPAS Achievement facet correlated most strongly with AIM Work Orientation (.36), indicating that it measures similar behaviors. TAPAS Non-Delinquency correlated with AIM Dependability (.46) and Hostility to Authority (-.44); all of these scales were intended to measure rule following and compliance with societal norms. As expected, TAPAS Dominance correlated .50 with the AIM and RBI Leadership scales, while showing much lower correlations with all other scales. Similarly, TAPAS Physical Conditioning correlated highly with AIM Physical Conditioning (.60) and RBI Fitness Motivation (.62) and much lower with everything else. Other TAPAS facets also showed predictable patterns of correlations with comparable personality scales from the AIM and RBI, indicating that TAPAS-95s measured the constructs it was intended to measure.

Table 4.11. Correlations between TAPAS-95s Facets and Selected Dimensions from the AIM, RBI, and ASVAB

Inventory	Dimension	TAPAS-95s Facet											
		ACH	CUR	NDLQ	DOM	EVTE	ATTN	I.E.	ORD	PHYC	TOL	COOP	OPT
AIM	Adjustment	.13	.20	.16	.05	.32	-.17	.13	.00	.09	.12	-.03	**.39**
	Agreeableness	.09	.16	.26	-.04	**.40**	-.25	.05	.00	.05	.07	.07	.19
	Dependability	.16	.16	**.46**	.10	.15	-.31	.07	.11	.00	.06	-.02	.07
	Leadership	.19	.22	.03	**.50**	.02	.05	.23	.05	.10	.13	-.24	.06
	Physical Conditioning	.22	.10	.00	.04	.06	-.06	.02	.05	**.60**	.06	-.12	.05
	Work Orientation	**.36**	.23	.05	.22	.12	-.08	.17	.09	.30	.12	-.23	.08
	Lie Scale	.00	-.06	-.02	-.04	.01	-.02	-.03	-.02	.02	-.03	-.03	.02
RBI	Leadership	.15	.22	.03	**.42**	.04	.08	.23	.02	.13	.17	-.19	.06
	Cognitive Flexibility	.13	**.41**	.09	.17	.17	-.09	**.33**	-.03	.03	.25	-.09	.08
	Achievement	**.23**	.19	.19	.22	.01	-.06	.14	.11	.13	.13	-.13	-.02
	Fitness Motivation	.18	.06	-.09	.08	.04	.02	.06	-.02	**.62**	.02	-.18	.08
	Stress Tolerance	.16	.17	.06	.08	**.26**	-.12	.20	-.01	.14	.09	-.07	**.31**
	Hostility to Authority	-.14	-.18	**-.44**	-.06	-.19	**.34**	-.09	-.08	.05	-.08	-.06	-.10
ASVAB	AFQT	-.06	**.24**	.06	.06	.14	-.07	**.38**	-.04	.00	.02	-.04	.18

Note. N = 2,422 – 3,362. ACH = Achievement; CUR = Curiosity; NDLQ = Non-Delinquency; DOM = Dominance; EVTE = Even Tempered; ATTN = Attention Seeking; I.E. = Intellectual Efficiency; ORD = Order; PHYC = Physical Conditioning; TOL = Tolerance; COOP = Cooperation; OPT = Optimism.

EEEM Criterion Validity Results

Detailed criterion results for TAPAS-95s are presented in the EEEM research technical report (Knapp & Heffner, 2010). Consequently, the focus here is on incremental validities of the TAPAS-95s facets relative to the cognitive ability composite (AFQT), which is currently used for selection and classification. Table 4.12 shows such results for 8 selected criteria that were measured at the end of their Initial Entry Training (either Advanced Individual Training [AIT] or

One-Station Unit Training [OSUT]). TAPAS-95s was administered at the beginning of basic training, so the study was longitudinal.

As can be seen in Table 4.12 considerable incremental validities were observed for various adjustment, graduation, and attrition criteria. For example, AFQT had a .05 correlation with 6-month attrition, but adding a composite of TAPAS-95s facets increased the overall multiple regression coefficient, R, to .24. This increase could be considered substantial given the multiplicity of reasons attrition occurs. The highest observed overall R for the AFQT and TAPAS-95s facets was for the Adjustment to Army Life criterion (.36), followed by the BCT graduation (.31), and the Last Army Physical Fitness Test (.31). These results clearly showed that including a personality inventory in the U.S. Army selection and classification test battery could help to identify applicants who are more motivated to finish their training and are capable of meeting the physical and emotional demands of military life.

Table 4.12. Incremental Validity Results for TAPAS-95s Facets and Eight Training Criteria

Criterion	*Incremental Validity*			
	N	AFQT Only	AFQT + TAPAS-95s	Δ *R*
Adjustment to Army Life Scale (ALQ)	523	.13	.36	.23
Last Army Physical Fitness Test (APFT) Score (ALQ; Self-Reported)	522	.04	.30	.26
Number of Disciplinary Incidents (ALQ; Self-Reported)	523	.11	.27	.17
Comprehensive Graduate vs. Discharged from Training (Reception through AIT/OSUT)	1,237	.03	.23	.20
Four-Month Attrition	1,694	.03	.27	.24
Six-Month Attrition	1,694	.05	.24	.19
AIT/OSUT - Graduate vs. Discharged	990	.00	.31	.31
Average Technical Training Exam Scores	585	.14	.23	.10

Note. Δ*R* = Increment in multiple correlation. Nagelkerke's *R* was used for the dichotomous criterion variables (Comprehensive Graduate vs. Discharged from Training; Four-Month Attrition, Six-Month Attrition, AIT/OSUT - Graduate vs. Discharged).

EEEM Adverse Impact Results

The final set of results for the first field testing of TAPAS concerns ethnic and gender subgroup differences. Because this assessment system was intended mainly for use in personnel selection and classification contexts, the presence of marked differences in scale means across these groups (a.k.a. adverse impact) could limit test use. Table 4.13 shows TAPAS and AFQT scale comparisons for males vs. females (M-F, Column 2), Blacks vs. Whites (B-W, Column 3), and Hispanics vs. White, non-Hispanics (H-WNH, Column 4). In each comparison, negative values indicate lower means for protected groups. To facilitate interpretation, standardized group differences are reported.

As can be seen from Table 4.13, TAPAS scales showed predictable patterns of gender differences. Males had somewhat higher Physical Conditioning, Optimism, and Even Tempered scores, while females had somewhat higher Non-Delinquency, Order, and Tolerance scores. The magnitudes of the differences were small, with none exceeding .30 in either direction. For the ethnic differences, adverse impact results were even more encouraging. While the AFQT showed standardized mean differences of -.63 and -.51, none of the TAPAS-95s scales exhibited differences larger than .25 in any direction. In fact, the largest score differences in Black vs. White comparisons were in favor of Blacks, who were more dominant, tolerant and orderly. Hispanics also were more tolerant than Whites, but their scores on Intellectually Efficiency and Non-Delinquency were somewhat lower.

In sum, subgroup comparisons for TAPAS facets revealed little if any impact against members of protected groups. In fact, minorities and women earned higher TAPAS scores than members of comparison groups on several scales, meaning that if the TAPAS scores were used in conjunction with AFQT for selection decisions, the overall impact against protected groups would be reduced.

Table 4.13. Subgroup Comparisons of AFQT and TAPAS-95s Scores

	Gender Differences	Race/Ethnic Differences	
Predictor	F-M *d*	B-W *d*	H-WNH *d*
AFQT	-0.30	-0.63	-0.51
TAPAS			
Achievement	0.12	-0.06	-0.07
Curiosity	0.09	0.08	0.01
Non-Delinquency	0.30	-0.01	-0.16
Dominance	0.29	0.21	-0.02
Even Tempered	-0.16	-0.01	-0.09
Attention Seeking	0.00	-0.04	0.04
Intellectual Efficiency	-0.15	0.04	-0.17
Order	0.25	0.21	0.02
Physical Cond.	-0.26	0.03	0.03
Tolerance	0.25	0.25	0.13
Cooperation	0.09	-0.05	-0.01
Optimism	-0.15	-0.04	0.07

Note. N = 2,422. F-M = female vs. male; B-W = Black vs. White; H-WNH = Hispanic vs. White-non-Hispanic.

Summary and Conclusions

In this chapter, we presented the results of three lines of work designed to investigate the validity of TAPAS. In the first, a meta-analysis was conducted to identify the facets and broad factors of personality that predict task, contextual, and counterproductive aspects of performance in military, civilian police, and civilian firefighter jobs. Forty-three studies starting in the year 1988 satisfied our inclusion criteria. Correlations were corrected for sampling error and unreliability. The results showed that personality variables were most useful for predicting aspects of contextual and counterproductive performance and, as was expected, cognitive aptitude measures were generally more effective for predicting successful task performance. Importantly, the results provide an empirical basis for using TAPAS facets to predict outcomes in the Armed Services. An important byproduct of this work is a database comprising over 1600 correlations that can help users choose TAPAS facets for specific testing applications.

In the second line of research described in this chapter, we examined the comparability of MDPP, single statement, and unidimensional pairwise preference personality tests by administering these measures under controlled conditions. The results provided strong evidence of score equivalence when respondents have little motivation to distort their answers. Facet scores from the three formats not only showed high correlations, but also nearly identical criterion validities, thus providing strong support for the measurement approach used in TAPAS.

Finally, the longitudinal EEEM research involved the development and administration of the first MDPP test constructed from the TAPAS statement pool. The test, called TAPAS-95s, was administered to a large sample of Army recruits at the beginning of training, along with a wide variety of self- and other- source outcome measures, which were collected over a period of several months. Statistical analyses indicated that TAPAS-95s scores correlated as expected with related scales in the AIM and RBI, providing further evidence of construct validity. Importantly, unit weighted composites of TAPAS-95s scores provided substantial incremental validity relative to the AFQT for many performance outcomes. In addition, TAPAS-95s scores showed little evidence of adverse impact against protected groups: Score differences in all subgroup comparisons were small, and the directions of differences varied to produce no appreciable net effect.

Together, these three areas of research provided strong support for the conceptual and empirical roots of TAPAS and made a cogent case for proceeding with the development and exploration of computerized adaptive testing (CAT) for operational selection environments.

CHAPTER 5: COMPUTERIZED ADAPTIVE TESTING WITH MULTIDIMENSIONAL PAIRWISE PREFERENCE ITEMS

This chapter describes the development of a computer adaptive testing algorithm for administering MDPP tests. We discuss the advantages of CAT with pairwise preference items and present an example of an information surface, which provides values needed for efficient item selection. We then present the results of an extensive simulation study that compared estimation accuracy for adaptive and nonadaptive MDPP tests involving varying numbers of dimensions (3, 5, 7, 10), items per dimension (5, 10, 20), and proportions of unidimensional pairings (5, 10, 20) needed to fix the score scale. Results indicated that adaptive MDPP item selection provided improvements in scoring accuracy relative to nonadaptive item selection, similar to what has typically been observed with unidimensional single stimulus IRT models. As in Chapter 1, the term "dimensions" is used here in place of "facets"; and items per dimension provides a convenient basis for comparing scoring accuracy with tests of different dimensionality and length. For example, a 5-d test involving 8 items per dimension would consist of 40 pairwise preference items, whereas a 10-d test involving 8 items per dimension would comprise 80 pairwise preference items.

Introduction to MDPP Adaptive Testing

Pairwise preference items are attractive for noncognitive assessment because they seem to retain the benefits of more complex forced choice formats, such as resistance to response sets (Borman et al., 2001; Brown & Maydue-Olivares, 2011), while being simpler to answer. They are also more tractable from a mathematical modeling standpoint, which is important considering the goal of adaptive testing. Even with constraints on how statements representing various dimensions are paired, based on content, extremity, and social desirability specifications, a pool of 500 statements measuring 13 dimensions can realistically yield tens of thousands of unique MDPP items. Permuting design specifications across examinees or altering test specifications over time can further enhance test security through reduced statement and item exposure and, coupled with the possibilities that model-based measurement provides for detecting aberrant response patterns, measurement based on MDPP items provides a solid psychometric basis for field applications.

With pairwise preference items that involve statements representing different dimensions, the relationship between trait levels and endorsement probabilities cannot be represented simply by a trace line, but instead requires a three-dimensional surface. Item response surfaces are somewhat difficult to describe, because they exhibit a number of peaks and valleys, but they directly relate trait levels on the dimensions represented by the respective statements in a pair to the probability of preferring one statement to the other. An example item response surface involving personality statements representing Dominance and Responsibility is shown in Figure 5.1.

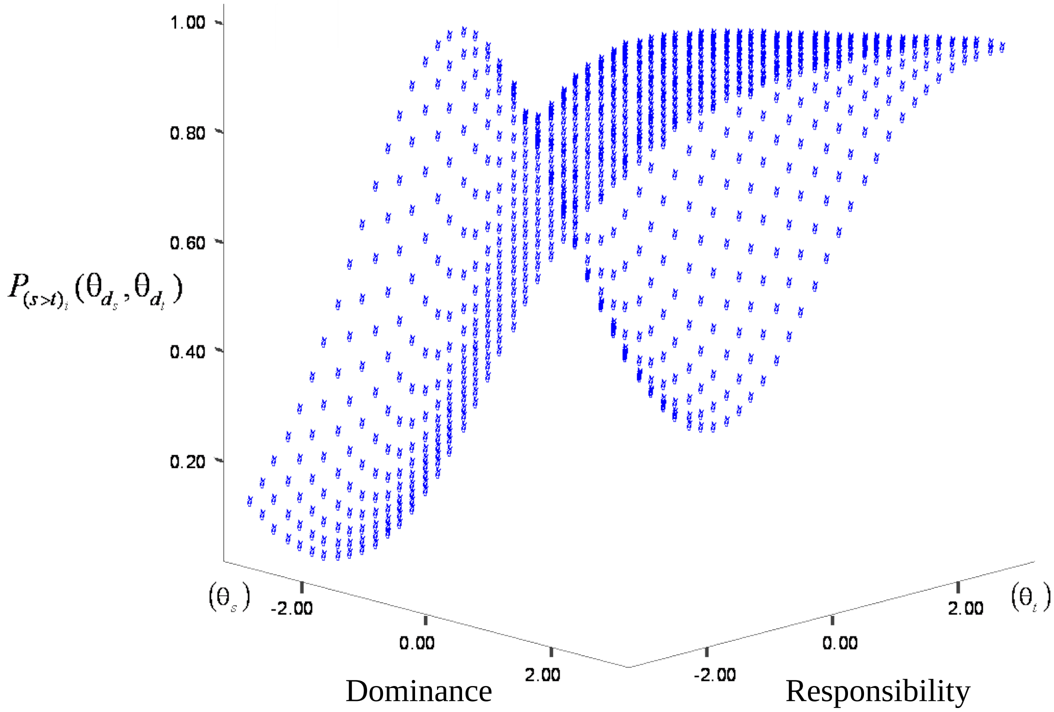

Figure 5. 1. Example item response surface for a multidimensional pairwise preference item measuring Dominance and Responsibility.

In Figure 5.1, values along the vertical axis indicate the probability of preferring statement s to statement t given a respondent’s standing on the respective dimensions and each statement’s GGUM parameters; these values were computed using the preferential choice probability equation described in Chapter 1 of this report. Importantly, at any combination of trait levels, we can also compute the amount of information provided by a pairwise preference item using equations shown in Stark et al. (2005). The relationship between trait levels and item information can be illustrated graphically using an item information surface, an example of which is shown in Figure 5.2. This surface can be interpreted in the same manner as traditional information functions. Information is inversely related to the error of measurement, so where information is higher, error is lower, and vice versa. In adaptive testing with MDPP, the goal is to construct items by selecting pairs of statements so that they are highly informative about the respondent's standings on the traits assessed, given the current estimates of his or her trait values. In this way, it is possible to substantially reduce the number of items required for accurate trait estimation.

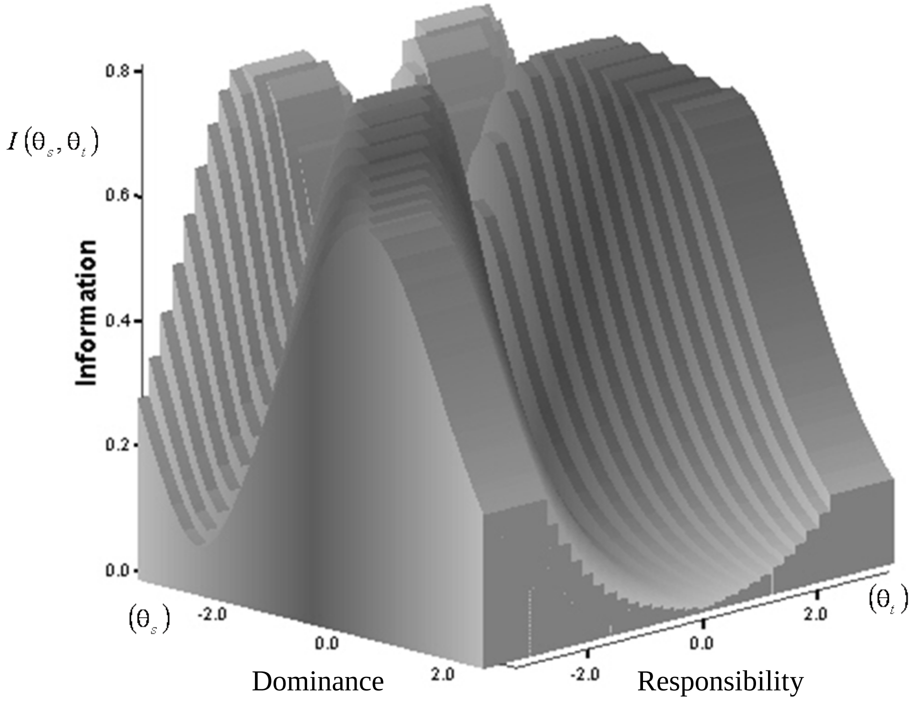

Figure 5.2. Example MUPP item information surface for a multidimensional pairwise preference item.

The CAT Algorithm

Adaptive MDPP testing in applied settings must address three issues. First, one must determine the dimensions that will be assessed and develop pools of statements that vary adequately in terms of location and, in the case of personality testing, social desirability. (This process for TAPAS was described in Chapter 3.) Second, constraints must be implemented to pair statements in a way that will not only identify the scoring metric, but will also enhance resistance of the test to faking. Third, one must decide whether to terminate testing based on estimated standard errors of trait scores or based on a fixed number of items. With both administration time and perceived fairness in mind, we developed a "fixed length" adaptive algorithm, which is described below. The algorithm was written in Visual Basic .NET.

1. Specify the number of dimensions to assess and the number of *items per dimension*. These choices determine the total test length. For example, one might choose to assess 15 dimensions with 8 items per dimension, and thus create an assessment involving 120 pairwise preference items.
2. Create and store content codes representing all permissible multidimensional and unidimensional combinations (e.g., for a 3-d test, 1-1, 2-2, 3-3, 1-2, 1-3, 2-3). This is

accomplished by identifying all possible combinations of dimensions and then excluding those that have been ruled out for substantive reasons.

3. To allow estimation of trait scores as soon as possible during a test, assume a respondent has an initial trait score of zero (the prior mean) on each dimension, administer a subset of items based on a circular linking design (e.g., for a 5-D test, present item types 1-2, 2-3,3-4,4-5, along with one unidimensional pair), estimate the respondent's trait scores using a multidimensional maximization procedure (e.g., Stark, 2002), and then continue sampling item types heterogeneously from the domain of permissible combinations subject to information, content, location, and social desirability constraints. (An alternative to this automated test design method is to have a test administrator set the sequence of pairs of dimensions a priori to increase consistency across examinees. Such is the case with the current versions of TAPAS administered in the MEPS.)
4. After each item is administered, estimate the respondent's trait scores and continue testing until the designated number of items has been reached. After the last item has been administered, compute the final trait scores and standard error estimates and save the results.

A Simulation Study Comparing the Effectiveness of MDPP CAT and Nonadaptive Tests

A VB.NET program was used to compare the efficacy of the adaptive algorithm, described above, with nonadaptive tests of the same length and dimensionality. Estimation accuracy was examined using a fully crossed design involving different numbers of dimensions (3, 5, 7, 10, 25), items per dimension (5, 10, 20), percentages of unidimensional pairings (5, 10, 20), and correlations between dimensions (.0, .3, .5) to examine the accuracy of scores and, moreover, evaluate the robustness of the scoring algorithm to violations of the assumption that the latent traits are independent. In each condition, data were generated by sampling 1,000 trait scores from multivariate standard normal distributions, and trait score recovery was assessed using correlations with generating parameters and bias statistics. Note that to increase the realism of the simulations, we utilized statement parameter estimates and social desirability ratings from the actual TAPAS pool. However, some minor changes were made to improve the balance of discrimination and location values across dimensions.

Results

Table 5.1 presents the correlations between estimated and generating trait scores, and Table 5.2 presents the average absolute error statistics for the nonadaptive and adaptive test simulations. In each table, the first column shows the correlation between the generating thetas ($\square_{\text{gen}}$) sampled for each dimension. The second column shows the percentage of items that were unidimensional. The third column indicates the number of items per dimension; for example, a 3-D test involving five items per dimension would comprise 15 items. The remaining columns show the results for the nonadaptive and adaptive conditions respectively. In each case, the value shown in a cell represents the average of the statistic across dimensions.

Table 5.1. Comparison of Correlations between Estimated and Known Trait Scores for Nonadaptive and Adaptive MDPP Tests

Average Correlation Across Dimensions												
			Nonadaptive					Adaptive				
ρ_{gen}	% Unidim.	Items Per Dimension	3-D	5-D	7-D	10-D	25-D	3-D	5-D	7-D	10-D	25-D
0	5	5	.72	.68	.72	.73	.75	.84	.83	.85	.84	.84
		10	.83	.82	.84	.84	.85	.90	.90	.90	.90	.89
		20	.91	.90	.91	.92	.92	.94	.94	.94	.94	.94
	10	5	.71	.68	.72	.72	.75	.85	.84	.85	.84	.83
		10	.83	.82	.84	.84	.84	.90	.90	.90	.90	.89
		20	.91	.90	.91	.92	.92	.93	.93	.94	.94	.95
	20	5	.71	.69	.70	.71	.73	.85	.84	.84	.84	.84
		10	.82	.80	.83	.83	.84	.90	.90	.91	.91	.89
		20	.90	.90	.91	.91	.92	.94	.93	.94	.94	.95
.3	5	5	.69	.66	.69	.70	.74	.82	.81	.82	.81	.80
		10	.82	.79	.81	.83	.84	.89	.89	.89	.89	.87
		20	.91	.89	.91	.92	.92	.93	.93	.93	.94	.94
	10	5	.68	.67	.68	.70	.73	.89	.82	.82	.82	.80
		10	.83	.80	.82	.83	.83	.89	.90	.90	.89	.87
		20	.90	.89	.90	.91	.91	.93	.93	.91	.94	.94
	20	5	.69	.65	.66	.69	.72	.83	.83	.83	.82	.81
		10	.81	.79	.80	.82	.83	.90	.90	.90	.90	.88
		20	.90	.88	.89	.91	.91	.93	.94	.94	.94	.94
.5	5	5	.66	.64	.65	.69	.72	.81	.80	.79	.79	.77
		10	.81	.78	.81	.82	.83	.88	.88	.88	.88	.85
		20	.90	.88	.90	.91	.91	.92	.92	.93	.93	.93
	10	5	.65	.63	.66	.67	.71	.88	.80	.81	.80	.78
		10	.79	.79	.80	.81	.82	.88	.88	.88	.89	.86
		20	.90	.89	.90	.91	.91	.92	.93	.93	.93	.94
	20	5	.64	.63	.65	.67	.70	.82	.82	.82	.81	.80
		10	.80	.78	.79	.82	.82	.90	.89	.89	.89	.87
		20	.90	.89	.89	.90	.91	.93	.93	.93	.94	.94

Note. The results are based on samples of N=1,000 simulated examinees in all but the “25-D, 20 items per dimension” conditions. In those cells, only 300 examinees were run due to very long computing times.

As can be seen in Table 5.1, the average correlations between estimated and known trait scores increased as test length increased from 5 to 20 items per dimension, and there was virtually no effect for the percent of unidimensional pairings. These results are consistent with the 1-D and 2-D results reported by Stark et al. (2005). Moreover, the fact that 5% unidimensional pairings was adequate for attaining good rank order recovery of trait scores, regardless of the number of dimensions assessed, is important from an applied perspective, because unidimensional items are arguably less resistant to faking than multidimensional items and therefore we wish to minimize their use. Finally, note the striking improvements in the correlations between estimated and generating trait scores when going from nonadaptive to adaptive item selection. Adaptive tests yielded approximately the same correlations as nonadaptive tests that were nearly twice as long. In conditions where the generating thetas were correlated $\rho_{gen} = .3$ and $\rho_{gen} = .5$, only minor decreases in estimation accuracy were observed, despite the sharp contrast with the assumptions of the independent normal prior distributions used for trait estimation. Importantly, the overall pattern of results was the same and the average correlations were virtually the same for the adaptive tests involving anywhere from 3 to 25 dimensions.

Table 5.2 presents the average absolute bias statistics

$$\frac{1}{DN}\sum_{d=1}^{D}\sum_{j=1}^{N}\left|\hat{\theta}_{dj}-\theta_{dj}\right|$$

for the nonadaptive and adaptive tests. Here $\hat{\theta}_{dj}$ is the estimated trait value for simulee *j* on dimension *d* for *N* simulees responding to a *D* dimensional test.

The findings shown in Table 5.2 mirror those for the correlations shown in Table 5.1. Specifically, there appears to be no effect for the percent of unidimensional pairings, which suggests that 5%, rather than 10%, would be sufficient for fixing the scale. Second, as before, test length had the primary effect on estimation accuracy, with longer tests showing smaller absolute biases (as tests were lengthened, regression to the mean effects were reduced). Also, as before, adaptive tests performed considerably better than their nonadaptive counterparts, requiring only half as many items to achieve the same level of accuracy. Third, the average absolute bias in estimated trait scores was almost identical for similar types of tests involving as many as 25 dimensions. This suggests that the estimated trait scores are not only useful for rank ordering examinees for selection applications, but also in terms of assessing overall accuracy.

Table 5.2. Comparison of Absolute Bias for Nonadaptive and Adaptive MDPP Tests

Average Absolute Bias Across Dimensions												
			Nonadaptive					Adaptive				
ρ_{gen}	**% Unidim.**	**Items Per Dimension**	**3-D**	**5-D**	**7-D**	**10-D**	**25-D**	**3-D**	**5-D**	**7-D**	**10-D**	**25-D**
0	**5**	**5**	.54	.57	.54	.53	.52	.42	.42	.41	.42	.42
		10	.43	.44	.42	.41	.41	.33	.33	.33	.33	.35
		20	.33	.33	.31	.30	.29	.27	.26	.26	.25	.25
	10	**5**	.55	.57	.54	.53	.52	.41	.41	.41	.42	.42
		10	.43	.44	.42	.42	.41	.33	.33	.33	.32	.35
		20	.32	.33	.32	.31	.30	.27	.27	.25	.25	.24
	20	**5**	.54	.56	.56	.54	.53	.41	.41	.42	.42	.42
		10	.44	.45	.44	.42	.42	.33	.32	.32	.32	.34
		20	.33	.34	.32	.32	.31	.27	.27	.26	.25	.24
.3	**5**	**5**	.57	.57	.55	.55	.53	.44	.43	.43	.45	.47
		10	.44	.46	.44	.43	.42	.35	.33	.34	.35	.38
		20	.32	.34	.32	.31	.30	.28	.27	.27	.27	.26
	10	**5**	.56	.57	.56	.55	.54	.43	.43	.43	.44	.47
		10	.44	.46	.43	.43	.43	.34	.33	.34	.34	.37
		20	.33	.34	.33	.32	.31	.28	.26	.31	.26	.26
	20	**5**	.56	.57	.57	.56	.54	.43	.42	.42	.44	.45
		10	.45	.46	.45	.44	.43	.34	.33	.33	.34	.37
		20	.34	.35	.34	.33	.31	.28	.26	.26	.26	.26
.5	**5**	**5**	.57	.59	.58	.56	.53	.46	.45	.44	.47	.49
		10	.45	.48	.45	.44	.43	.36	.36	.36	.36	.39
		20	.34	.35	.33	.32	.32	.29	.29	.28	.28	.28
	10	**5**	.59	.59	.58	.57	.54	.44	.45	.45	.46	.48
		10	.46	.46	.46	.45	.43	.36	.35	.35	.35	.39
		20	.33	.35	.33	.32	.31	.29	.28	.28	.28	.27
	20	**5**	.60	.59	.59	.57	.55	.44	.44	.44	.45	.46
		10	.46	.47	.46	.44	.44	.33	.34	.35	.35	.37
		20	.34	.35	.34	.33	.32	.28	.28	.27	.27	.27

Note. The results are based on samples of N=1,000 simulated examinees in all but the “25-D, 20 items per dimension” conditions. In those cells, only 300 examinees were run due to long computing times.

The main shortcoming identified in this simulation dealt with difficulties in estimating standard errors for the observed trait scores. The Bayes modal method used to estimate trait scores and standard errors for the MUPP model is based on the D-dimensional minimization/maximization subroutine DFPMIN described by Press, Flannery, Teukolsky, and Vetterling (1990) that utilizes a Broyden-Fletcher-Goldfarb-Shanno estimation algorithm. Press et al. suggested that this algorithm requires as many iterations as there are parameters estimated to produce accurate standard errors via the approximated inverse Hessian. Because DFPMIN is called sequentially in this application and fewer internal iterations occur when the likelihood function is near a maximum, the standard errors obtained in this manner tend to be considerably larger than the standard deviations computed over replications for the same set of generating trait scores (empirical standard deviations), particularly when the number of parameters estimated is large (Stark & Drasgow, 2002). Here, this implies that the estimated standard errors will increase as dimensionality increases, even if the trait scores are estimated with similar accuracy.

This phenomenon is illustrated in Table 5.3 below, which presents the estimated standard errors for the trait scores having the correlations and absolute bias statistics shown in Tables 5.1 and 5.2.

Table 5.3. Comparison of Estimated Standard Errors for Nonadaptive and Adaptive MDPP Tests Based on the Inverse Hessian Approximation

Average Estimated Standard Error Across Dimensions using Inverse Hessian												
			Nonadaptive					**Adaptive**				
ρ_{gen}	**% Unidim.**	**Items Per Dimension**	**3-D**	**5-D**	**7-D**	**10-D**	**25-D**	**3-D**	**5-D**	**7-D**	**10-D**	**25-D**
0	**5**	**5**	.65	.68	.71	.77	.88	.47	.44	.46	.53	.76
		10	.52	.53	.58	.67	.83	.38	.36	.38	.45	.68
		20	.40	.40	.45	.58	.78	.32	.26	.33	.43	.70
	10	**5**	.66	.68	.72	.77	.88	.47	.44	.46	.53	.76
		10	.53	.54	.58	.68	.83	.38	.36	.38	.46	.69
		20	.40	.40	.45	.58	.80	.32	.31	.33	.44	.70
	20	**5**	.67	.69	.73	.78	.89	.47	.42	.46	.54	.77
		10	.53	.54	.59	.68	.84	.38	.36	.38	.46	.70
		20	.40	.41	.46	.59	.80	.32	.27	.33	.44	.70
.3	**5**	**5**	.69	.67	.71	.77	.88	.47	.43	.45	.52	.76
		10	.52	.53	.57	.67	.83	.37	.35	.37	.44	.67
		20	.39	.40	.44	.57	.79	.31	.30	.33	.42	.68
	10	**5**	.66	.68	.72	.77	.88	.46	.43	.45	.52	.76
		10	.52	.53	.58	.67	.83	.37	.35	.37	.44	.68
		20	.40	.40	.45	.59	.80	.31	.30	.36	.43	.70
	20	**5**	.67	.69	.73	.78	.89	.46	.42	.45	.53	.77
		10	.53	.54	.59	.68	.84	.37	.35	.37	.46	.69
		20	.40	.41	.46	.59	.81	.31	.30	.32	.43	.71
.5	**5**	**5**	.66	.68	.71	.77	.88	.46	.43	.44	.51	.75
		10	.52	.53	.57	.67	.83	.36	.35	.36	.44	.67
		20	.39	.40	.45	.58	.80	.29	.30	.32	.43	.70
	10	**5**	.66	.68	.72	.77	.88	.45	.42	.44	.52	.76
		10	.52	.53	.58	.67	.83	.36	.35	.36	.44	.67
		20	.39	.40	.45	.58	.80	.31	.29	.32	.44	.70
	20	**5**	.66	.69	.73	.78	.89	.45	.42	.44	.52	.77
		10	.53	.54	.59	.68	.84	.36	.34	.37	.46	.69
		20	.40	.41	.46	.59	.81	.31	.29	.32	.45	.71

Note. The results are based on samples of N=1,000 simulated examinees in all but the “25-D, 20 items per dimension” conditions. In those cells, only 300 examinees were run due to long computing times.

As can be seen in the Table 5.3, the estimated standard errors increased as test dimensionality increased, rising most sharply from the 10-D to the 25-D conditions. This is in contrast to the stable correlations and bias statistics, suggesting that the increase was largely an artifact attributable to decreasing quality of the inverse Hessian approximation. To deal with this issue, a replication method for obtaining standard error estimates analogous to empirical standard deviations was proposed and tested for a representative set of conditions in the following investigation.

Examining Standard Errors Estimated using a New Replication Method

One way to benchmark the accuracy of standard error estimation is to compute the standard deviation of trait score estimates over replications on a set of grid points reflecting a range of true trait scores. Stark and Drasgow (2002), for example, examined standard error estimation via the DFPMIN routine by simulating response data for 100 examinees having trait scores of -3.0, -2.8, …, +3.0 on a unidimensional grid and comparing the standard errors estimated via the inverse Hessian approximation to the standard deviations of the estimated thetas over replications.

With that idea in mind, we developed a replication method for estimating standard errors of TAPAS trait scores. Upon conclusion of a test, the final trait scores are used to generate multiple response patterns based on the parameters for the statements composing the pairwise preference items that were administered. One hundred replications would be desirable, but infeasible due to long computing times, so 30 replications were used for this investigation; 50 are used currently in TAPAS. Data were simulated according to the MUPP model and the standard deviations of the respective trait score estimates over replications were computed to assess the variation in the scores due to random error. Greater variation was expected for trait scores derived using nonadaptive tests than those from adaptive tests, because the adaptive trait score estimates were usually more accurate due to increased test information.

To compare the stability and relative magnitude of the replication-based standard errors to those obtained via the inverse Hessian approximation, we conducted a simulation involving 300 examinees sampled from multivariate normal distributions for nonadaptive and adaptive tests of 3 to 25 dimensions and correlations among generating trait scores ranging from .0 to .5. 100-item fixed length tests, with 10% of the items being unidimensional, were chosen to provide a smaller, but representative set of study conditions. Correlations between estimated and known trait scores (Corr), standard errors estimated via the inverse Hessian (EstSE) and replication (SErep) methods, bias, and absolute bias (AbsBias) results are presented in Table 5.4.

Table 5.4. Trait Score Recovery Statistics Showing Estimated and Replication-Based Standard Errors for 100-Item MDPP Tests Involving 10% Unidimensional Pairings

		Nonadaptive					Adaptive				
ρ_{gen}	**Average Across Dimensions**	**3-D**	**5-D**	**7-D**	**10-D**	**25-D**	**3-D**	**5-D**	**7-D**	**10-D**	**25-D**
0	**Corr**	.84	.81	.84	.84	.84	.90	.89	.90	.91	.89
	EstSE	.52	.54	.58	.68	.81	.38	.36	.38	.45	.67
	SErep	.42	.43	.42	.41	.40	.31	.32	.32	.32	.33
	Bias	-.04	-.03	-.01	-.04	-.04	.00	.05	.01	.00	-.01
	AbsBias	.43	.44	.43	.42	.41	.33	.34	.33	.32	.35
.3	**Corr**	.81	.80	.81	.82	.83	.91	.89	.90	.90	.87
	EstSE	.52	.53	.58	.68	.82	.37	.35	.37	.45	.66
	SErep	.41	.43	.42	.42	.40	.31	.32	.31	.32	.33
	Bias	-.04	-.09	-.06	-.05	-.08	-.01	.00	.00	.02	-.01
	AbsBias	.45	.47	.45	.43	.42	.32	.35	.34	.33	.38
.5	**Corr**	.78	.78	.78	.81	.82	.86	.88	.89	.89	.86
	EstSE	.51	.54	.58	.68	.81	.36	.35	.37	.45	.66
	SErep	.41	.43	.42	.42	.40	.31	.31	.31	.31	.32
	Bias	.04	-.05	-.07	-.07	-.07	.03	.01	.01	.00	.01
	AbsBias	.47	.48	.48	.45	.43	.35	.36	.36	.36	.39

Note. Corr = correlation between estimated and generating trait scores. EstSE = standard errors estimated via the approximated inverse Hessian matrix. SE rep = standard errors computed using replication method. AbsBias = absolute bias of estimated trait scores. The results are based on samples of N=300 simulated examinees due to long computing times.

As can be seen in the table, the correlations and bias results changed very little as test dimensionality increased, even when the correlations between generating trait values increased from .0 to .5. The SErep values also were very stable and, as expected, about .1 smaller for the adaptive tests than for the comparable nonadaptive tests regardless of dimensionality. In contrast, the EstSE values increased by about 60% under the same conditions, suggesting that the values were spuriously high due to too few iterations in the parameter estimation routine when convergence occurred. For this reason, TAPAS uses the replication method to estimate standard errors of trait estimates.

Summary and Conclusions

In high volume assessment settings, there is a growing predilection for test construction and delivery methods that increase efficiency by reducing test length and readily providing scores for decision making. Historically, adaptive testing applications have been limited to the cognitive ability and academic achievement domains. The research presented here, however, describes the development and evaluation of an adaptive testing methodology designed specifically for noncognitive assessment involving multidimensional pairwise preference items. Such items belong to the general class of forced choice measures that are becoming increasingly popular in organizational settings because they seem to provide more resistance to response sets than single statement measures.

In this chapter, we presented the results of an extensive simulation investigation which showed that MDPP trait scores could be recovered with good to excellent accuracy for relatively short tests involving up to 25 dimensions. It was found, for example, that with a 10-D test involving just 100 items, five of which were unidimensional, the average correlations across dimensions between generating (known) and estimated trait scores were .84 and .90 for nonadaptive and adaptive tests, respectively, and adaptive tests performed about as well as nonadaptive tests that were nearly twice as long. Moreover, when we examined scoring accuracy using generating trait scores correlated .3 with each other, thus violating assumptions of the independent normal priors used for Bayes modal estimation, the correlations between observed and generating trait scores dropped by just .01, suggesting that correlations between facets of the same broad factors would not limit applications to idealized simulation conditions. Together these results strongly supported the test construction and scoring approach underlying TAPAS.

CHAPTER 6: IMPLEMENTATION OF TAPAS TESTING AT MILITARY ENTRANCE PROCESSING STATIONS

In 2009, the U.S. Army approved the initial operational testing and evaluation (IOT&E) of the TAPAS for use with Army applicants at military entrance processing stations (MEPS). Dimensions comprising the MEPS version of TAPAS were selected on the basis of our meta-analytic results (reported in Chapter 4), with the long term goal of creating personality composites that might be used in conjunction with cognitive measures to improve selection and classification decisions. This testing pool contained over 800 personality statements, and was thus large enough to generate tens of thousands of pairwise preference items tailored to the trait levels of individual applicants for enlistment. Statement parameters for this pool were estimated from data collected from large samples of new recruits from 2006 to 2008 (see Chapter 3). For the actual computer adaptive administration, items were selected dynamically using an algorithm like the one described in Chapter 5.

Three versions of TAPAS were created for MEPS testing using this initial pool. Because detailed results comparing those versions are presented in an ARI technical report for the Army Tier One Performance Screen (TOPS) IOT&E (see Knapp, Heffner, & White, 2011), this chapter focuses primarily on results relevant to the construct validity of the 15-dimension TAPAS CAT that was administered in MEPS until July 2011, when it was replaced by a second statement pool. The composition of the current MEPS TAPAS and the steps required to create new TAPAS measures are presented in the next chapter.

Description of TAPAS Testing at MEPS

Three computerized versions of TAPAS were developed for MEPS testing. The first version was a 13-dimension computerized adaptive test (CAT) containing 104 pairwise preference items. This version is referred to as the TAPAS-13D-CAT. TAPAS-13D-CAT was administered from May 4, 2009 to July 10, 2009 to about 2,200 recruits. In July 2009, TAPAS MEPS testing was expanded to 15 dimensions by adding the facets of Adjustment from the Emotional Stability domain and Self Control from the Conscientiousness domain, and test length was increased to 120 items. In both cases, testing time was limited to 30 minutes.

Two 15-dimension TAPAS tests were created. One version was nonadaptive, so all examinees answered the same sequence of items; the other was adaptive, so each examinee answered items tailored to his/her trait level estimates. The TAPAS-15D-Static was administered from mid-July to mid-September of 2009 to all examinees, and thereafter continuously to smaller numbers of examinees at some MEPS (these MEPS were slow to replace the static version with the adaptive version). The adaptive version, referred to as TAPAS-15D-CAT, was introduced in September of 2009 and was administered to a large number of recruits until July 2011 when it was replaced by a newer version based on a second item pool. Table 6.1 shows the facets assessed by the 13-dimension and 15-dimension measures.

Table 6.1. TAPAS Facets Assessed at MEPS

Facet Name	Brief Description	Big Five Factor
Dominance	High scoring individuals are domineering, “take charge” and are often referred to by their peers as "natural leaders."	Extraversion
Sociability	High scoring individuals tend to seek out and initiate social interactions.	
Physical Conditioning	High scoring individuals tend to engage in activities to maintain their physical fitness and are more likely participate in vigorous sports or exercise.	
Attention Seeking	High scoring individuals tend to engage in behaviors that attract social attention. They are loud, loquacious, entertaining, and even boastful.	
Selflessness	High scoring individuals are generous with their time and resources.	Agreeable-ness
Cooperation	High scoring individuals are pleasant, trusting, cordial, non-critical, and easy to get along with.	
Achievement	High scoring individuals are seen as hard working, ambitious, confident, and resourceful.	Conscientiousness
Order	High scoring individuals tend to organize tasks and activities and desire to maintain neat and clean surroundings.	
Self Control[a]	High scoring individuals tend to be cautious, levelheaded, able to delay gratification, and patient.	
Non-Delinquency	High scoring individuals tend to comply with rules, customs, norms, and expectations, and they tend not to challenge authority.	
Adjustment[a]	High scoring individuals are well adjusted, worry free, and handle stress well.	Emotional Stability
Even Tempered	High scoring individuals tend to be calm and stable. They don’t often exhibit anger, hostility, or aggression.	
Optimism	High scoring individuals have a positive outlook on life and tend to experience joy and a sense of well-being.	
Intellectual Efficiency	High scoring individuals believe they process information and make decisions quickly; they see themselves (and they may be perceived by others) as knowledgeable, astute, or intellectual.	Openness To Experience
Tolerance	High scoring individuals are interested in other cultures and opinions that may differ from their own. They are willing to adapt to novel environments and situations.	

Note. [a] Not included in TAPAS-13D-CAT.

The administration procedures for the three TAPAS versions were identical. Each testing session was initiated by a test administrator who entered the examinee's identifying information. Next, each examinee was asked to read information related to the purpose of the assessment. Army applicants were told that the test scores would be used for selection purposes, and Air Force applicants were apprised that the scores would be used for research purposes only. Then an instruction page appeared. This page provided detailed information about answering TAPAS items and showed some examples. Examinees were told to consider how they typically think, feel, and act, and to indicate which statement in each pair (i.e., each pairwise preference item) was "more like me." They were informed that some pairs would be difficult to answer and, in such cases, they should consider both options carefully and indicate the one that described them, perhaps just slightly, better than the other. After making their choice by clicking on the appropriate statement, they should affirm their response and continue with the assessment by clicking the "Next Item" button. Testing proceeded in this manner until all items were completed or the 30 minute time limit elapsed. Scores were considered "valid" only if an examinee completed at least 80% of the items. (Note that in the event of a test interruption, the administrator could save the session and restart the assessment at the same point.)

Detailed results for each testing session were saved and transferred to a central database upon test completion. These include item responses and response time for each item, trait scores, the number of minutes taken to complete the entire test, flags to detect fast responders, and two composites known as Can Do and Will Do that were developed in the EEEM research **(**Allen et al., 2010; Knapp & Heffner, 2010**)**. The Can Do composite was created to predict criteria such as MOS-specific job knowledge, AIT exam grades, and graduation from AIT/OSUT. The Will Do composite was designed to predict criteria such as physical fitness, adjustment to Army life, effort, and support for peers.

Construct Validity Results for TAPAS MEPS Testing

Table 6.2 shows means, standard deviations, and intercorrelations for the 15 personality facets measured by the TAPAS-15D-CAT, along with AIM and AFQT scores. (For more details and results for the TAPAS-13D-CAT and TAPAS-15D-Static tests, please see Knapp et al., 2011.) The correlation matrix in Table 6.2 was obtained from a sample of 120,356 U.S. Army applicants tested between July, 2009 and February, 2011 at 65 MEPS in the continental U.S. and Hawaii. This sample was predominantly male (80.7%). A majority of participants (65%) had a high school diploma or its equivalent, with another 17.3 % still in high school. 7.5 % completed university degrees, 3.6% were still in college and another 3.2% had associate degrees. 65.4% were White, followed by 15.2% Hispanic and 11.8% African American. 18.6% did not indicate their racial background. The average AFQT score for the sample was 56.2 (SD = 24.3), with 58.7% obtaining an AFQT score of 50 or higher.

The results in Table 6.2 showed that the operational TAPAS exhibited patterns of correlations consistent with expectations for the AIM dimensions, the AFQT composite, and the Big Five factors with which they are associated. For example, Intellectual Efficiency, which is a facet of Openness to Experience, correlated .41 with the AFQT composite. This is consistent with previous studies that have found Openness to Experience and to be positively correlated with measures of achievement and aptitude (Ashton, Lee, Vernon, & Jang, 2000; Higgins, Peterson, Pihl, & Lee, 2007) and what was found in the EEEM research (see Chapter 4). Correlations between the more closely connected personality dimensions of TAPAS and AIM were also similar to those found in the EEEM study. For example, Dominance correlated .43 with AIM Leadership in the current investigation and .51 in the EEEM research. Achievement correlated .38 with AIM Work Orientation in the present investigation and .34 in the EEEM research, while Non-delinquency correlated .34 with AIM Dependability in the present investigation and .46 in the EEEM study. Intercorrelations among TAPAS dimensions belonging to the same broad personality factor were also in the expected directions. For example, the No Anxiety, Even Tempered, and Well Being dimensions, which are facets of the Big Five factor Emotional Stability, correlated in the .18 to .30 range, while Dominance, Sociability and Attention Seeking, all facets of Extraversion, correlated in the .22 to .35 range. The similarity in findings across the present investigation and the EEEM research and the pattern of correlations of TAPAS facets in the present investigation suggest that constructs being measured by TAPAS under operational conditions are highly similar to those that were measured in a research setting. We expect, therefore, that criterion related validities for TAPAS assessments in operational conditions will also be similar to those observed in research settings.

Table 6.2. Correlations between 15-D 120-Item TAPAS, AIM, and AFQT

	Dimension	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	Achievement	.15	.48																					
2	Adjustment	-.01	.57	.10																				
3	Cooperation	-.07	.38	.11	.12																			
4	Dominance	.03	.57	.32	.10	.01																		
5	Even Tempered	.15	.47	.10	.18	.24	-.05																	
6	Attention Seeking	-.20	.52	.04	.11	.05	.20	-.01																
7	Selflessness	-.19	.43	.09	-.02	.21	.01	.10	-.07															
8	Intellectual Efficiency	-.03	.58	.26	.19	.04	.25	.09	.08	-.01														
9	Non-delinquency	.08	.46	.17	.00	.16	-.01	.18	-.13	.12	.02													
10	Order	-.40	.54	.15	-.08	.00	.04	-.03	-.08	.05	.02	.07												
11	Physical Conditioning	.03	.61	.15	.07	-.01	.18	-.07	.12	-.04	.05	-.02	.03											
12	Self Control	.05	.54	.21	.06	.11	.05	.20	-.12	.06	.17	.25	.15	-.05										
13	Sociability	-.04	.58	.05	.11	.18	.22	.04	.35	.07	.01	-.04	-.04	.13	-.10									
14	Tolerance	-.22	.56	.11	.03	.14	.06	.12	.04	.31	.07	.05	.04	-.06	.10	.11								
15	Optimism	.13	.45	.19	.26	.16	.17	.19	.16	.04	.11	.09	-.02	.10	.07	.23	.09							
16	AIM: Dependability	1.33	.21	.17	-.02	.08	-.01	.16	-.06	.12	.04	.34	.06	-.05	.19	-.02	.08	.07						
17	AIM: Adjustment	1.35	.22	.15	.23	.10	.04	.18	.12	.01	.17	.12	-.04	.02	.10	.10	.06	.22	.46					
18	AIM: Work Orientation	1.36	.22	.38	.06	-.01	.23	.06	.12	.09	.21	.08	.04	.07	.12	.08	.07	.09	.23	.37				
19	AIM: Leadership	1.24	.23	.22	.06	-.03	.43	-.05	.21	.03	.18	.01	.00	-.02	.03	.25	.07	.10	.18	.19	.43			
20	AIM: Agreeableness	1.38	.21	.08	.08	.20	-.06	.26	.06	.09	.05	.11	-.02	-.07	.13	.06	.09	.13	.37	.45	.27	.07		
21	AIM: Physical Conditioning	1.38	.32	.18	.01	.01	.17	-.01	.08	.00	.08	.05	.04	.38	.05	.01	.01	.06	.30	.35	.35	-.04	.21	
22	AFQT	56.17	24.30	.10	.11	.02	.09	.09	.11	-.06	.41	.00	-.19	.05	.00	-.08	-.01	.04	.08	.18	.17	.09	.06	.01

Note. AFQT = U.S. Armed Forces Qualifying Test. AIM = Assessment of Individual Motivation temperament questionnaire. Sample size for AFQT and TAPAS was 120,356; sample size for AIM was 3,259. Sample includes applicants for Regular Army, U. S. Army National Guard, and U. S. Army Reserve.

Additional evidence for the construct validity of the TAPAS-15D-CAT is provided in Table 6.3, which shows means and standard deviations of TAPAS scores for male (N = 97,165) and female (N = 23,170) applicants whose scores were used to compute the correlations among TAPAS facets shown in Table 6.2. (Note the small difference in sample sizes between two tables, which is due to missing demographic data). The MEPS results are very similar to those that were observed in the EEEM research (Knapp & Heffner, 2010). Males had higher scores on Adjustment, Intellectual Efficiency, and Physical Conditioning, whereas females scored higher on Tolerance, Selflessness, Order, and Non-Delinquency. Importantly, the gender differences were relatively small in magnitude and did not favor either group consistently. Consequently, they would essentially offset each other when composites are formed.

Table 6.3. Female-Male Comparisons of TAPAS Scale Scores among U.S. Army Applicants at MEPS

	Females		Males		F - M
TAPAS Facet	Mean	SD	Mean	SD	d
Achievement	0.17	0.46	0.15	0.48	0.04
Adjustment	-0.14	0.56	0.02	0.57	-0.29
Cooperation	-0.08	0.37	-0.06	0.38	-0.03
Dominance	-0.02	0.56	0.05	0.58	-0.12
Even Tempered	0.11	0.47	0.16	0.46	-0.11
Attention Seeking	-0.24	0.51	-0.19	0.52	-0.11
Selflessness	-0.06	0.43	-0.23	0.43	0.37
Intellectual Efficiency	-0.12	0.54	0.00	0.59	-0.21
Non-delinquency	0.13	0.44	0.06	0.46	0.15
Order	-0.33	0.55	-0.41	0.53	0.15
Physical Conditioning	-0.16	0.59	0.08	0.61	-0.40
Self Control	0.05	0.54	0.05	0.53	0.00
Sociability	-0.03	0.58	-0.04	0.58	0.02
Tolerance	-0.07	0.53	-0.25	0.56	0.34
Optimism	0.12	0.46	0.13	0.45	-0.04

Note. F = Female (N = 23,170); M = Male (N = 97,165); d = mean difference (F-M). Sample includes applicants for Regular Army, U. S. Army National Guard, and U. S. Army Reserve..

Comparisons of Operational and Research Only TAPAS Administrations

Among the most interesting and important results for the TAPAS-15D-CAT are the scale score comparisons of Army and Air Force applicants who took the test for different purposes. Recall that Army applicants took TAPAS under operational testing instructions and their scores were used for selection; thus, they had a clear incentive to do well. In contrast, Air Force applicants were told that TAPAS scores would be used for research purposes only, so they had no reason to distort their answers in an effort to raise scores. This situation might therefore be seen as a naturally occurring field experiment that allows us to examine the susceptibility of TAPAS to faking.

Table 6.4 presents facet means, standard deviations, and standardized differences for Army and Air Force applicants who took TAPAS-15D-CAT between July 2009 and May 2011. To maximize the comparability of the groups, these samples included only Regular Army and Air Force applicants and excluded those applying for the National Guard or Reserves. As can be seen from the table, Army and Air Force applicants showed remarkably similar scores on all dimensions. In fact, none of the 15 comparisons revealed a difference exceeding .15 standard deviations. Army applicants scored slightly higher on Adjustment and Dominance, whereas Air Force applicants scored slightly higher on Non-Delinquency and Even Tempered. Overall, these results indicate that TAPAS scores showed virtually no differences across operational and research conditions.

Table 6.4. Descriptive Statistics for TAPAS CAT Scores in Regular Army and Air Force Samples

	Army		Air Force		Army - Air Force
TAPAS Facet	Mean	SD	Mean	SD	d
Achievement	0.16	0.49	0.13	0.50	0.07
Adjustment	0.01	0.57	-0.04	0.58	0.08
Cooperation	-0.07	0.38	-0.04	0.38	-0.07
Dominance	0.03	0.57	-0.05	0.59	0.13
Even Tempered	0.15	0.46	0.19	0.46	-0.08
Attention Seeking	-0.20	0.52	-0.19	0.52	-0.01
Selflessness	-0.21	0.44	-0.22	0.45	0.02
Intellectual Efficiency	-0.02	0.59	0.01	0.60	-0.05
Non-delinquency	0.07	0.47	0.14	0.46	-0.14
Order	-0.40	0.54	-0.43	0.56	0.06
Physical Conditioning	0.03	0.61	0.06	0.64	-0.04
Self Control	0.05	0.54	0.03	0.54	0.04
Sociability	-0.05	0.58	-0.06	0.59	0.01
Tolerance	-0.21	0.55	-0.24	0.56	0.05
Optimism	0.12	0.45	0.14	0.45	-0.04

Note. Sample Sizes: Regular Army = 86,962; Air Force = 30,658; Examinees with unusual response latencies (e.g., very fast) and unusual statement selections were removed prior to analyses.

Summary

This chapter described the composition of 13-D and 15-D computerized TAPAS forms that were administered in MEPS from October 2009 through early 2011. Results showing the relationships between TAPAS-15D-CAT, AIM, and AFQT scores were presented. TAPAS scores were shown to correlate predictably with relevant AIM dimensions, and the Intellectual Efficiency facet of TAPAS correlated positively, as anticipated, with the AFQT, providing evidence of convergent and discriminant validity. We also presented results showing the similarity of TAPAS-15D-CAT across large samples of males and females. Score differences were very small and in different directions, suggesting that composites formed from TAPAS scores will be gender neutral. In addition, we presented means and standard deviations of TAPAS scores collected from Army and Air Force applicants who took the test for very different purposes. Army applicants took TAPAS for operational decision making and thus had incentives to do well. Air Force applicants took TAPAS for research purposes only and had no incentives to fake good. The results of this naturally occurring field experiment unequivocally showed that Army and Air Force applicants had similar scores on all dimensions. Although it would be premature to draw definitive conclusions about the susceptibility of TAPAS to faking, or other types of response distortion, these results provide the strongest evidence to date that the MDPP methodology underlying TAPAS is viable for high stakes testing applications.

CHAPTER 7: DEVELOPMENT OF A SECOND TAPAS STATEMENT POOL AND IMPLEMENTION OF NEW VERSIONS OF TAPAS CAT AT MEPS: A GUIDE FOR CREATING NEW TAPAS CONSTRUCTS

Following the implementation of TAPAS testing in MEPS, Drasgow Consulting Group was funded to develop another statement pool for exclusive use by the Department of Defense. The primary reason was to enhance test security: Subsets of statements in the original pool had been used in other military applications. A second but no less important purpose was to improve and expand the initial pool by adding five new facets: Situational Awareness, Team Orientation, Courage, Adventure Seeking, and Commitment to Serve, all of which lie outside of the Big Five framework, but were deemed relevant to military selection and classification.

This chapter describes the development of a second statement pool measuring the original 22 TAPAS facets plus the five new facets that were requested by ARI. This pool was implemented with an updated version of the TAPAS CAT executable, which randomly chooses one of three test configurations when an applicant is tested at a MEPS. Each version of the TAPAS CAT measures 15 dimensions. Nine dimensions are common across versions to provide some comparability and to compute the Can Do and Will Do composites described earlier. Six dimensions are unique to each version to allow collection of validity data on most of the remaining facets. The layouts of the test versions are illustrated later in this chapter. The statements in the new pool, along with their psychometric properties (GGUM parameters and social desirability ratings), are provided in a separate report.[1]

Descriptions of the Five New TAPAS Facets

Five new facets were added to the TAPAS taxonomy in an effort to enhance the capabilities of the assessment system. The five facets tapped into behavioral domains directly relevant to military life and, thus, had a potential to increase the validity of TAPAS composites for selection and classification decisions. The first new facet is Adventure Seeking, which focuses primarily on high intensity, high risk outdoor activities. Individuals scoring high on this facet enjoy participating in extreme sports and outdoor activities and might consider typical forms of activity boring. It was expected that this facet would be particularly relevant for predicting performance and retention of Soldiers requiring long periods of outdoor activity, such as Infantry and Special Forces military occupational specialties.

The second new facet is Commitment to Serve, which assesses one’s level of identification with the military and commitment to a military lifestyle. High scoring individuals respect the military and take pride in being able to serve their country. This facet was included in the Rational Biographical Inventory (RBI; Kilcullen, Putka, McCloy, & Van Iddekinge, 2005) and has shown promise in predicting Soldier retention.

[1] This report has limited distribution with approval of ARI to provide item security.

The third new facet is Courage, which assesses how brave and daring applicants are when faced with adversity. High scoring individuals stand up to challenges and indicate a willingness to operate in dangerous situations. This facet is expected to predict combat performance, retention, and re-enlistment intentions.

Team Orientation is the fourth new TAPAS facet. Behaviors associated with this facet deal with the desire to work in a team environment. This facet is expected to predict peer and supervisory performance ratings of teamwork, especially in jobs requiring extensive group activities.

The final new facet is Situational Awareness. Scores reflect how vigilant and attentive Soldiers are to their external environments. Individuals scoring high on this facet pay attention to their surroundings and rarely get lost or surprised. This facet was deemed particularly relevant for predicting performance in combat and guard-related duties.

Table 7.1 provides descriptions of the five facets that were added to the TAPAS taxonomy along with examples of statements representing high and low levels of the constructs.

Table 7.1. Description of Five New TAPAS Facets

New TAPAS Facet	Brief Description of High Scorer
Adventure Seeking	High scoring individuals enjoy participating in extreme sports and outdoor activities.
Commitment to Serve	High scoring individuals identify with the military and have a strong desire to serve their country.
Courage	High scoring individuals stand up to challenges and are not afraid to face dangerous situations.
Team Orientation	High scoring individuals prefer working in teams and make people work together better.
Situational Awareness	High scoring individuals pay attention to their surroundings and rarely get lost or surprised.

Development of the Second Statement Pool

The development of the second statement pool proceeded in a manner similar to the TAPAS item development described in Chapter 3 of this report. First, we wrote 50-70 statements for each original TAPAS facet and the five new facets. The statements were written to evenly span the respective trait continua. This was verified by having two SMEs (PhD faculty members in Industrial/Organizational Psychology) rate all of the statements on a scale of 1 (low) to 7 (high). Statements showing obvious discrepancies across raters were discarded or revised, and gaps in the distributions were addressed by writing additional statements to produce a fairly even distribution for each facet. Because five new facets were added to the TAPAS taxonomy and some of the original 22 facets were not being considered as candidates for MEPS testing, a decision was made by ARI to drop four from the original taxonomy, Aesthetics, Depth, Ingenuity and Virtue, leaving 23 facets for which pretest data were needed for parameter estimation. Next, the statements were carefully edited for grammar, punctuation, and readability.

At this point, all newly written statements for the 23 facets were reviewed again for length, clarity, sensitivity, and, for the 19 original facets, the degree of content overlap with the original pool. Some statements were modified slightly to improve readability and some were flagged for removal by ARI due to high similarity. Statements that passed this review were assembled into pretest forms and administered to the samples of Army recruits shown in Table 7.2. All data collections complied with the Army's and the American Psychological Association ethical guidelines for research with human subjects.

Table 7.2. Samples Used to Estimate GGUM Parameters for the Second TAPAS Statement Pool

Location	Date	Sample Size
Fort Leonard Wood	10-Aug-09	528
Fort Leonard Wood	18-Aug-09	462
Fort Jackson	16-Oct-09	524
Fort Benning	9-Jul-10	837
Fort Leonard Wood	1-Aug-10	789
Fort Sill	15-Aug-10	1302
Fort Leonard Wood	22-Aug-10	778
Total		5220

Pretest Questionnaire Construction

For each testing session, a pretest questionnaire with multiple forms was developed. Multiple forms were needed to efficiently collect the data required for estimating GGUM statement parameters and social desirabilities. (GGUM estimation requires 400 to 500 persons per statement, whereas fairly stable social desirability estimates can be obtained using samples of just 50.) Common subsets of 5 to 7 statements per facet were included in questionnaire forms administered within and across testing sessions so that parameter estimates could be placed on a common scale.

Each form of a questionnaire contained two sections. The first section required examinees to respond honestly. The second section required examinees to fake good. In both sections, data were collected using a four-point response format, where 1 = strongly disagree, 2 = disagree, 3 = agree, and 4 = strongly agree. The honest section always preceded the faking section in a questionnaire because it was believed that it would be easier for examinees to shift from an honest to a faking mindset than the reverse.

The honest section of each questionnaire form contained 180-220 statements measuring six to ten facets of personality. Respondents were instructed to honestly and accurately indicate their level of agreement with each statement using the four-point (strongly disagree to strongly agree) format. For illustration, the instruction page for the honest section of a pretest questionnaire is presented in Appendix A.

The faking section of each questionnaire form contained 40 to 50 statements reflecting varying numbers of dimensions. In this section, respondents were told to pretend they were not yet in the Army, but very much wanted to be and scores on the test that followed would be used to make admissions decisions. Thus, they should answer in a way that made them look like "good Army material." As in the honest section, responses were collected using a four-point format. For illustration, the instruction page for the faking section of a pretest questionnaire is presented in Appendix B.

Procedures Followed During Testing Sessions

Typically 90 to 160 recruits participated in each testing session. When possible, the testing materials were distributed before the recruits were brought into the testing room. Anywhere from one to nine forms of a pretest questionnaire were administered in each session, but each recruit completed just one form. When multiple forms of a questionnaire were administered in a single session, the booklets were spiraled to distribute them evenly across the room in order to obtain approximately the same number of responses to each form (Kolen & Brennan, 2004).

Each testing session began with the proctors introducing themselves and referring recruits to the Privacy Act and Informed Consent form that had been distributed along with testing materials and an index card containing a unique randomly assigned identification number. The nature and purpose of testing was summarized and recruits were told to read the consent form carefully, and if they agreed to participate, they should sign and date the form. Next, examinees were asked to complete the background information on the optical scanning forms that would be used to record their responses. The random number on each index card was entered on the forms to link the sheets for each person while maintaining anonymity. The proctor then asked examinees to open the pretest questionnaire and turn to the instructions for section one, which required honest responding. The proctor carefully reviewed and explained the instructions and sample item, emphasizing the importance of answering honestly and accurately, and examinees were told that they would have approximately 50 minutes to complete the section. In addition, they were instructed to sit quietly and not to proceed with the next section if they finished early.

Upon completion of the honest section, the recruits were given a brief break. Participants were then told that a short but important task remained. They were asked to open their questionnaires to section two (the faking section) and pay close attention as the instructions were read aloud. The instructions and an example statement were reviewed and the importance of answering in a way that maximized the chances of being accepted into the Army were strongly emphasized. Participants were given approximately 20 minutes to complete this short section, and most finished in under 15.

Data Analysis

Data from the recruit samples were processed and cleaned to remove invalid entries. Data from the honest conditions were dichotomized and analyzed for each dimension separately, using the GGUM2000 computer program (Roberts, 2001). Three GGUM parameters were estimated for each statement: discrimination (α), location (δ), and threshold (τ). The GGUMLINK computer program (Roberts, 2002) was used to put the new statements on the original TAPAS scale based on the common subsets of five to seven statements that were systematically included in the various forms. The polytomous data from the faking conditions were then used to estimate one social desirability parameter per statement by averaging the endorsed response codes over examinees.

In summary, over 1200 new statements measuring 18 of the original TAPAS facets and 5 new facets were developed. These statements were written to reflect low, medium, and high locations on each trait continuum. These statements were pretested on large samples of Army recruits in four different U.S. Army installations over a period of about one year. GGUM and social desirability parameters were estimated for each statement. Statements having GGUM discrimination parameters below .50 were then eliminated, because they would have been very unlikely candidates for inclusion in MDPP items. At the same time, ARI statements, which had been used in the 2009 TAPAS MEPS testing, were moved into the second pool, because they had not been exposed in TAPAS-related testing outside of the MEPS. In total, this effort produced 1142 usable statements for the second TAPAS pool, with the final numbers for each facet as shown in Table 7.3.

Table 7.3. Numbers of Statements Representing each of the 23 Facets in the Second TAPAS Statement Pool

Facet Name	Number of ARI Statements	Number of New Statements	Total
Original Facets			
Achievement	11	46	57
Adjustment	9	44	53
Aesthetics		discontinued	discontinued
Attention Seeking		47	47
Consideration		56	56
Cooperation	8	38	46
Curiosity		46	46
Depth		discontinued	discontinued
Dominance	14	37	51
Even Tempered	9	44	53
Ingenuity		discontinued	discontinued
Intellectual Efficiency		44	44
Non-Delinquency	12	34	46
Optimism	8	39	47
Order		50	50
Physical Conditioning	19	35	54
Responsibility		42	42
Self Control		42	42
Selflessness		56	56
Sociability		48	48
Tolerance		44	44
Virtue		discontinued	discontinued
New Facets			
Adventure Seeking		51	51
Commitment to Serve		52	52
Courage		56	56
Situational Awareness		48	48
Team Orientation		53	53
Total	90	1052	1142

Note. “discontinued” indicates that the facet was not chosen for inclusion in the pool.

2011 TAPAS Testing at MEPS using the Second Pool

Twenty-one of the available 23 facets were selected for 2011 MEPS testing based on the second pool. Sixteen facets were chosen from the original group of 18; these included the 15 facets that appeared in the 2009 TAPAS-15D-CAT as well as Responsibility. This latter facet was expected to enhance the prediction of contextual performance, counterproductivity, and leadership effectiveness. The remaining five facets were the newly developed ones (Adventure Seeking, Commitment to Serve, Courage, Situational Awareness, and Team Orientation), which targeted behavioral domains particularly relevant for military jobs.

A new TAPAS CAT executable program was developed in VB.NET and installed at the MEPS. This program was designed to randomly choose one of three TAPAS versions to administer to an examinee upon the initiation of a new testing session by a MEPS proctor. The versions varied in composition to ensure that each of the selected 21 facets would be administered to at least 40% of examinees overall. As with the 2009 implementation, each version of TAPAS CAT was designed to assess just 15 facets using 120 pairwise preference items, and testing time was set to a maximum of 30 minutes.

To uniquely identify data collected using the new TAPAS versions, and in light of the three previous computerized versions that were administered in MEPS between October 2009 and early 2011, versions of TAPAS based on the new second pool were numbered beginning with 5. In the early development stages, four configurations were proposed and created, but one was dropped to mitigate administration and data management concerns identified by ARI. Consequently, the new TAPAS CAT configurations were labeled *Version 5*, *Version 7*, and *Version 8*.

Version 5 measures the same 15 facets that were assessed by the 2009 TAPAS-15D-CAT. It is set to be administered to 20% percent of MEPS examinees to provide a basis for judging the similarity of the new and original pools. We hope to see very similar patterns of means and intercorrelations for the old and new assessments, but some differences are anticipated due to refinements of the content domains, differences in statement wording, and, possibly, examinee characteristics.

Version 7 assesses 12 of the 15 facets that appeared in TAPAS-15D-CAT, along with three new facets: Adventure Seeking, Commitment to Serve, and Situational Awareness. This version is designed to be administered to 40% of MEPS examinees.

Version 8 also contains 12 facets that appeared in TAPAS-15D-CAT; 9 of these 12 facets overlap with Version 7. In addition, Version 8 assesses Responsibility, which was previously excluded from MEPS testing, along with two new facets, Courage, and Team Orientation. Version 8 is also set to be administered to 40% of MEPS applicants.

In summary, the partially overlapping designs for the new TAPAS versions ensure that each of the selected 21 *facets* will be administered to at least 40% of applicants. Nine facets from the 2009 implementation will be administered to all applicants, so that Can Do and Will Do composites can still be computed for enlistment eligibility decision making. The six remaining facets from the 2009 implementation will be administered to 60% of applicants, and the last six facets, Responsibility, Team Orientation, Commitment to Serve, Courage, Adventure Seeking, and Situational Awareness, will be administered to 40% of applicants to obtain new validity data. The layouts of the 2011 MEPS TAPAS versions are shown in Table 7.4.

Table 7.4. 2011 TAPAS CAT Layouts

	TAPAS Facet	2011 TAPAS CAT Software		
		V5	V7	V8
1	Achievement	*	*	*
2	Adjustment	*	*	*
3	Attention Seeking	*	*	*
4	Dominance	*	*	*
5	Even Tempered	*	*	*
6	Intellectual Efficiency	*	*	*
7	Non-Delinquency	*	*	*
8	Optimism	*	*	*
9	Physical Conditioning	*	*	*
10	Cooperation	*	*	
11	Order	*	*	
12	Self-Control	*		*
13	Selflessness	*	*	
14	Sociability	*		*
15	Tolerance	*		*
16	Adventure Seeking		*	
17	Commitment to Serve		*	
18	Courage			*
19	Responsibility			*
20	Situational Awareness		*	
21	Team Orientation			*

Note. V5 (Version 5) will be administered to 20% of applicants; V7 (Version 7) will be administered to 40% of applicants; and V8 (Version 8) will be administered to the remaining 40% of applicants.

Summary

This chapter describes in detail how a new statement pool for TAPAS testing was developed. It began with definitions of five new facets that were requested by ARI to increase the relevance of TAPAS dimensions for prediction of performance in military service. Statements were written for the facets of the original taxonomy plus five new facets. Pretest questionnaires involving honest and faking sections were administered to recruits at multiple installations to collect data on the statements needed to estimate GGUM and social desirability parameters. After parameter estimation, the pool was screened to remove poorly discriminating statements and ARI statements, which had been included in the 2009 testing pool, were subsequently moved into the new pool for Department of Defense use. The result was 1142 statements measuring 23 facets of personality. Of these 23 facets, 21 were selected for administration in MEPS, starting in early 2011. Three test layouts were created and a VB.NET executable was developed to randomly select a version to administer to each examinee. Each version assesses 15 dimensions using 120 pairwise preference items that must be answered in 30 minutes or less. The partially overlapping layouts and designated frequencies of administration for these versions, referred to as Versions 5, 7, and 8, ensure that the selected 21 facets will be administered in at least 40% of testing sessions.

Importantly, the information presented in this chapter serves as a “how to” guide for developing new TAPAS constructs from start to finish. The process of creating and administering pretest questionnaires is described, and Appendices A and B show sample instruction pages for the honest and faking sections of an illustrative questionnaire form. The information presented here and in Chapter 3 indicates the software and processes used for estimating GGUM and social desirability parameters. Once statement parameters have been estimated, it is a simple matter to import the information into the input data file that is used by TAPAS CAT software for item selection and scoring. Before administering a test, however, the layout or design specifications must be set by an administrator with substantive and psychometric knowledge of the process. Tests should be designed with a mix of unidimensional and multidimensional pairings. The simulation results in Chapter 5 suggest that 10% unidimensional pairings is enough to adequately set a scale and produce normative scores. With adaptive item selection, the simulation results and the practical limits on testing time suggest that eight items *per dimension* (i.e., facet) is a reasonable number considering the tradeoff between scoring accuracy and testing time. It is with this rule of thumb in mind that the total test length for the15D TAPAS CAT was set at 120 items (15 dimensions x 8 items/dimension = 120 items) . The CAT algorithm used for tailoring items to examinee trait levels is described in Chapter 5, and the methodology for scoring the dichotomous responses to these MDPP tests is explained in detail in Chapter 1 as well as in Stark et al. (2005).

CHAPTER 8: SUMMARY AND CONCLUSIONS

This report traces the development of TAPAS from initial studies through implementation and validation. Multiple lines of research have been conducted to evaluate the psychometric theory underlying TAPAS, the accuracy of estimation procedures, and validity in research and operational settings.

Numerous simulation studies have been conducted to test TAPAS's psychometric theory. The results clearly show that, when its assumptions are satisfied, trait estimates are accurate and satisfactory standard errors can be obtained via a replication method. Moreover, simulations show that CAT is efficient for three to 25 dimensions, for uncorrelated and correlated traits, for circular linking, and with only a small percentage of unidimensional comparisons.

MDPP trait estimates from research participants were compared to unidimensional forced choice trait estimates, and single statement items constructed from the TAPAS statement pool. They were also compared to trait estimates obtained from other assessment instruments such as the AIM and RBI. Substantial evidence for the construct validity of the MDPP trait estimates resulted.

Finally, and perhaps most importantly, TAPAS was evaluated in an IOT&E. Under operational conditions, no evidence of score inflation was obtained. Early results from the IOT&E indicate that, as expected, the ASVAB predicts Can Do criteria with little incremental validity for TAPAS. On the other hand, ASVAB does little to predict Will Do criteria such as physical fitness and attrition, whereas TAPAS explains important amounts of variance on these criteria.

This work has important implications for the U.S. Army. Specifically, there have been repeated calls for the use of non-cognitive predictors in military selection and classification (e.g., Drasgow et al., 2006). Until recently, however, such predictors were not implemented because of concerns about faking good. An important advance has been the successful use of the AIM for selection into the U.S. Army. TAPAS builds on AIM's advances by implementing computer adaptive assessment of a comprehensive set of personal characteristics. It has shown resistance to faking and evidence of its ability to predict important aspects of performance in the U.S. Army should soon be available.

REFERENCES

Ashton, M. C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior, 19,* 289-303.

Ashton, M. C., Lee, K., & Paunonen, S. V. (2002) What is the central feature of extraversion?: Social attention versus reward sensitivity. *Journal of Personality and Social Psychology, 83*, 245-251.

Ashton MC, Lee K, Vernon PA, Jang KL. (2000). Fluid intelligence, crystallized intelligence, and the Openness/intellect factor. *Journal of Research in Personality, 34,* 197–207.

Avolio, B., & Bass, B. M. (2002). *Manual for the Multifactor Leadership Questionnaire*. Redwood City, CA: Mindgarden.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bono, J. E., & Judge, T. A. (2004). Personality and transformational and transactional leadership: A meta-analysis. *Journal of Applied Psychology, 89,* 901–910.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10,* 99–109.

Briggs, S. R. (1989). The optimal level of measurement for personality constructs. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 246-260). New York: Springer.

Brown, A. & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*, 460-502.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56,* 81-105.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (2nd ed., Vol. 1, pp. 687-731). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P., & Knapp, D. J. (2001). Exploring the Limits in Personnel Selection and Classification, New Jersey: Lawrence Erlbaum Associates.

Cattell, R. B. (1973). *Personality and mood by questionnaire*. Oxford, England: Jossey-Bass.

Chernyshenko, O.S., Stark, S., Chan, K.Y., Drasgow, F., & Williams, B.A. (2001). Fitting Item Response Theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523-562.

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.

Chernyshenko, O.S., & Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M.D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22*, 105-127.

Conn, S., & Reike, M. L. (Eds.). (1994). *The 16PF fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.

Costa, P. T., Jr., & McCrae, R. R. (1988). From catalog to classification: Murray's needs and the five-factor model. *Journal of Personality and Social Psychology, 55*, 258-265.

Costa, P. T., Jr., McCrae, R. R., & Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences, 12*, 887-898.

Cronbach, L. J. & Gleser, G. C. (1959). Interpretation of reliability and validity coefficients: Remarks on a paper by Lord. *Journal of Educational Psychology, 50*, 230-237.

Dalal, R. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90*, 1241–1255.

De la Torre, J., Ponsoda, V., Leenen, I., & Hontangas, P. (2011). Some extensions of the Multi-Unidimensional Pairwise Preference Model. Paper presented at the 26th annual conference for the Society of Industrial and Organizational Psychology. Chicago, IL.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Drasgow, F., Chernyshenko, O.S., & Stark, S. (2010). 75 years after Likert: Thurstone was right (focal article). *Industrial and Organizational Psychology, 3*, 465-476.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB).* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155-166.

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

Goldberg, L.R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34.

Goldberg, L. R. (1997). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In: I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe*, Vol. 7. The Netherlands: Tilburg University Press.

Gough, H. G. (1987). *The California Psychological Inventory administrators guide*. Palo Alto, CA: Consulting Psychologists Press.

Guion, R.M., & Gottier, R.F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18,* 135-164.

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small scale cheating. *International Journal of Testing, 9,* 283-309.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74,* 167-184.

Higgins, D.M., Peterson, J.B., Pihl, R.O., & Lee, A.G.M. (2007). Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. *Journal of Personality and Social Psychology, 93,* 298–319.

Hofstee, W. K., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology, 63,* 146-163.

Hogan, R. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial & organizational psychology* (Vol. 2, pp. 873-919, 2nd ed.). Palo Alto, CA: Consulting Psychologists Press.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.

Hough LM. (2003). Emerging trends and needs in personality research and practice: Beyond main effects. In M.R. Barrick and A. M. Ryan (Eds.), *Personality and work:*

Reconsidering the role of personality in organizations (pp. 289–325). San Francisco, CA: Jossey-Bass.

Hough, L. M., & Ones, D. S. (2002). The structure, measurement, validity, and use of personality variables in industrial work, and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology, Vol. 1* (pp. 233-277). Thousand Oaks, CA: Sage Publications.

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.), *Current problems and techniques in multivariate psychology: Proceedings of a conference honoring Professor Paul Horst* (pp. 23-32). Seattle: University of Washington.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87-102). New York: Plenum Press.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications*. New York: Wiley.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71*, 327-333.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Jackson, D. N. (1994). *Jackson personality inventory – Revised manual*. Port Huron, MI: Sigma Assessment Systems, Inc.

Jackson, D. N., Wrobleski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance, 13*, 371-388.

Judge, T. A. (2009). Core self-evaluations and work success. *Current Directions in Psychological Science, 18*, 58–62.

Judge, T. A., & Bono, J. E. (2000). Five factor model of personality and transformational leadership. *Journal of Applied Psychology, 85,* 751–765.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*, 765–780.

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.) (2010). *Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., Heffner, T. S., & White, L. (Eds.) (2011). *Tier One Performance Screen initial operational test and evaluation: Early results* (Technical Report 1283). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences .

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practices*. 2nd Edition. New York: Springer.

Kriedt, P. H., & Dawson, R. I. (1961). Response set and the prediction of clerical job performance. *Journal of Applied Psychology, 45,* 175-178.

Lucas, R.E., Diener, E., Grob, A., Suh, E.M., & Shao, L. (2000). Cross-cultural evidence for the fundamental features of extraversion. *Journal of Personality and Social Psychology, 79,* 452–468.

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods, 8*, 222-248.

McCrae, R. R. (1996). Social consequences of experiential openness. *Psychological Bulletin*, *120*, 323-337.

McCrae, R. R. & Costa, P. T. (1985). Comparison of EPI and psychoticism scales with measures of the five-factor model of personality. *Personality and Individual Differences, 6*, 587-597.

McCrae, R. R. & Costa P. T. (1997). Conceptions and correlates of openness to experience. In R. Hogan; J. A. Johnson; & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 825-847). San Diego, CA: Academic Press.

Meade, A.W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531-551.

Mershon, B., & Gorsuch, R. L. (1988). Number of factors in personality sphere: Does increase in factors increase predictability of real life criteria? *Journal of Personality and Social Psychology, 55,* 675-680.

Mischel, W. (1968). *Personality and assessment*. New York: Wiley.

Mischel, W. (1969). Continuity and change in personality. *American Psychologist, 24*, 1012-1018.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*, 252-283.

Norman, N. M. (1963). The cultural context of personality theory. In J. M. Wepman & R. W. Heine (Eds.), *Concepts of personality* (pp. 333-360). Hawthorne, NY: Aldine Publishing.

Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of behavior. *Journal of Personality and Social Psychology, 74,* 538-556.

Paunonen, S. V., & Jackson, D. N. (2000). What is beyond the Big Five? Plenty! *Journal of Personality, 68,* 821-835.

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.

Organ, D. W. *(*1988). *Organizational citizenship behavior.* Lexington, MA: Lexington Books.

Roberts, B. W. (2002). *Study Behavior Questionnaire*. Unpublished manuscript, Department of Psychology, University of Illinois at Urbana-Champaign.

Roberts, B. W., Bogg, T., Walton, K., Chernyshenko, O. S., & Stark, S. (2004). A lexical investigation of the lower-order structure of conscientiousness. *Journal of Research in Personality, 38,* 164-178.

Roberts, B., Chernyshenko, O.S., Stark, S., & Goldberg, L. (2005). The construct of conscientiousness: The convergence between lexical models and scales drawn from six major personality questionnaires. *Personnel Psychology, 58,* 103-139.

Roberts, J. S. (2001). GGUM2000: Estimation of parameters in the generalized graded unfolding model. *Applied Psychological Measurement, 25,* 38.

Roberts, J. S. (2002). GGUMLINK Version 1.0: USER'S MANUAL. Unpublished Manuscript. Georgia Institute of Technology.

Roberts, J. S., Fang, H., Cui, W., & Wang, Y. (2006). GGUM2004: A Windows-based program to estimate parameters of the generalized graded unfolding model. *Applied Psychological Measurement, 30,* 64-65.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general model for unfolding unidimensional polytomous responses using item response theory. *Applied Psychological Measurement, 24,* 2-32.

Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal, 38,* 555–572.

Rotundo, M., & Sackett, P. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology, 87*, 66–80.

Roznowski, M.R., & Hanisch, K.A. (1990). Building systematic heterogeneity into work attitudes and behavior measures, *Journal of Vocational Behavior, 36*, 361-375.

Sackett, P. (2002). The structure of counterproductive work behaviors: Dimensionality and relationships with facets of job performance. *International Journal of Selection and Assessment, 10*, 5–11.

Saucier, G. (1992). Benchmarks: Integrating affective and interpersonal circles with the Big Five personality factors. *Journal of Personality and Social Psychology, 62,* 1025-1035.

Saucier, G., & Ostendorf, F. (1999). Hierarchical subcomponents of the Big Five personality factors: A cross-language replication. *Journal of Personality and Social Psychology, 76*, 613-627.

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

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. (2005). *EAP scoring program for ZG model*. Unpublished manuscript. University of Illinois: Urbana-Champaign.

Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86,* 943-953.

Stark, S., Chernyshenko, O.S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: An application to the problem of faking in personality assessment. *Applied Psychological Measurement, 29,* 184-201.

Stark, S., Chernyshenko, O.S., & Drasgow, F. (2010). Tailored Adaptive Personality Assessment System (TAPAS-95s). In D. J. Knapp and T.S. Heffner, T. S. (Eds.). *Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* pp. 15-22 (Technical Report 1267*)*. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Stark, S., Chernyshenko, O.S., & Drasgow, F. (April 2011). *Adaptive testing with the multi-unidimensional pairwise preference model.* Paper presented at the annual conference of the National Council on Measurement in Education. New Orleans, LA.

Stark, S., Chernyshenko, O.S., Drasgow, F., & Williams, B.A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25-39.

Stark, S., Chernyshenko, O. S., Lee, W. C., Drasgow, White, L. A., & F., Young, M. C. (2011). Optimizing prediction of attrition with the U.S. Army's Assessment of Individual Motivation (AIM). *Military Psychology, 23,* 180 – 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.

Tellegen, A. (1982). *A brief manual for the Multidimensional Personality Questionnaire*. Unpublished manuscript, University of Minnesota.

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.

Trent, T., & Quenette, M. A. (1992). *Armed Services Applicant Profile (ASAP): Development and validation of operation forms* (TR-92-9). San Diego, CA: Navy Personnel Research and Development Center.

Vickers, R. R., Jr., Conway, T. L., & Hervig, L. K. (1990). Demonstration of replicable dimensions of health behaviors. *Preventive Medicine, 19*, 377-401.

White, L. A., Nord, R. D., Mael, F. A., & Young, M. C. (1993). *The Assessment of Background and Life Experiences (ABLE).* In T. Trent & J. H. Laurence (Eds.), Adaptability screening for the Armed Forces. Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

White, L. A., & Young, M. C. (1998). *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, 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 (Eds.), *Exploring the limits of personnel selection and classification* (pp. 525-558). Mahwah, NJ: Erlbaum.

Viswesvaran, C., & Ones, D. (2000). Measurement error in “Big Five Factors” personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60*, 224–235.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71- 87.

Zinnes, J. L., & Griggs, R.A. (1974). Probabilistic, multidimensional unfolding analysis. *Psychometrika, 39,* 327 – 350.

APPENDIX A
EXAMPLE OF PRE-TEST FORM FOR ESTIMATING GGUM PARAMETERS OF TAPAS STATEMENTS

Instructions:

This section of the questionnaire asks you to respond to a series of statements describing how you typically think, feel, or act. It is very important that you **respond to the statements honestly.**

Read each statement carefully and decide the extent to which you agree or disagree. Then accurately fill in the corresponding oval on the scantron form. **Please do not write or mark on this questionnaire** – just indicate your answers on the scantron.

Work at a fairly rapid pace. **And, remember that you are to answer honestly.**

Sample Item

		STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
1.	I enjoy being part of a team.	a	b	c	d

Mark “a” on your scantron – if you strongly disagree with the statement

Mark “b” on your scantron – if you disagree with the statement

Mark “c” on your scantron – if you agree with the statement

Mark “d” on your scantron – if you strongly agree with the statement

Please Note:

- There are no right and wrong answers. Just respond to the items honestly and accurately.
- In choosing an answer, consider your life in general and not only the last few weeks or months.

- Some items may be difficult to answer. In those cases, just think a bit longer and choose the answer that best describes you.

- Some items will appear similar. This is not designed to trick you, so there's no need to look back at your previous answers. Just continue moving forward, answering the items honestly and accurately.

- Several items will ask you to mark a specific answer on your scantron. Your answers are used to check that our scanning software is working properly. Please make sure to mark the requested oval.

(Oral instructions: Are there any questions? If not, please turn the page and begin answering the items.)

Please indicate your answer to each item by marking the appropriate oval on the scantron. Remember, it is important that you **answer honestly and accurately.**

		STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
1.	I usually make a noticeable contribution to group problem-solving tasks.	a	b	c	d
2.	I am generally pretty forgiving.	a	b	c	d

APPENDIX B
EXAMPLE OF PRE-TEST FORM FOR ESTIMATING SOCIAL DESIRABILITY PARAMETERS OF TAPAS STATEMENTS

Instructions For The Remaining Section

Unlike in the previous sections where you were instructed to answer items as honestly and accurately as possible, we now want you to PRETEND that you are not yet in the Army, but very much want to be. Imagine that a recruiter has asked you to complete a test to determine if you are GOOD ARMY MATERIAL. If you score well, you will be let into the Army. If you don't score well, you won't.

For all remaining sections, we want you to answer items in a way that will make you look good from the Army's standpoint. In other words, answer in a way that will give the Army the best possible impression of you to insure that you pass the test and get accepted. Convince the Army that you will make a good Soldier!

Instructions:

This section of the questionnaire asks you to respond to a series of statements describing how one might think, feel, or act. The format is the same as one you saw earlier.

Remember, you are now trying to **create the best possible impression of yourself** from the **Army's standpoint**; so your answers do not need to describe you accurately. Just answer in a way that you think will maximize your chances of getting accepted into the Army.

Please do not write or mark on this questionnaire – just indicate your answers on the scantron.

Work at a fairly rapid pace. **And, convince the Army that you will make a good Soldier!**

Sample Item:

		STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
1.	I enjoy being part of a team.	a	b	c	d

Mark “a” on your scantron – if you think “Strongly Disagree” makes you like good Army material

Mark “b” on your scantron – if you think “Disagree” makes you like good Army material

Mark “c” on your scantron – if you think “Agree” makes you like good Army material

Mark “d” on your scantron – if you think “Strongly Agree” makes you like good Army material

Please Note:

- Some items may be difficult to answer. In those cases, just think a bit longer and choose the answer that best serves your “goal” of getting into the Army.

- Some items will seem similar. This is not designed to trick you, so there’s no need to look back at your previous answers. Just continue moving forward, answering in a way that makes you look like good Army material.

Please indicate your answer to each item by marking the appropriate oval on the scantron. Remember, **answer in a way that makes you like good Army material!**

	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
1. I have great respect for our legal system and support it in any way I can.	a	b	c	d
2. In group projects, I give personal and team goals equal weight and consideration.	a	b	c	d